Probabilistic Models of Physical Reasoning

Kevin A Smith[1,2], Jessica B Hamrick[3], Adam N Sanborn[4], Peter W Battaglia[3], Tobias Gerstenberg[5], Tomer D Ullman[1,2,6], and Joshua B Tenenbaum[1,2]

[1]Department of Brain and Cognitive Sciences, MIT

[2]Center for Brains, Minds, and Machines

[3]DeepMind

[4]Department of Psychology, University of Warwick

[5]Department of Psychology, Stanford University

[6]Department of Psychology, Harvard University

Probabilistic Models of Physical Reasoning

In order to reason about and interact with the world around us, we must understand how it changes over time. Crucially, we consider not just one possible future, but a range of possible outcomes: we can tell when a ball *almost* knocks another into a goal (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Gerstenberg & Tenenbaum, 2016), when a tower of blocks is precariously stacked and might fall down (Battaglia, Hamrick, & Tenenbaum, 2013), or that we are unsure where an occluded object will reappear (Smith & Vul, 2015). This suggests that our internal models of the physical world are probabilistic, translating uncertainty about the world's state or dynamics into a distribution of beliefs over possible future outcomes or latent object properties.

In this chapter we demonstrate how techniques from probabilistic modeling can be used to explain the predictions and inferences people make when reasoning about physical systems. We first describe why physical reasoning is an interesting problem, and why a probabilistic framing is important for tackling it. We then lay out one theory of probabilistic physical reasoning – the *Intuitive Physics Engine* (IPE; Battaglia et al., 2013). We discuss how probabilistic modeling with the IPE can explain a wide range of ways people reason about physics. Next we describe how the mind might perform this reasoning efficiently, through approximations to both probabilistic reasoning and the IPE. We end with current and future directions for probabilistic models of physical reasoning.

The probabilistic nature of physical reasoning

In many other cases in this book, it is possible to exactly calculate the posterior probability distributions necessary for probabilistic reasoning. Indeed, classic work demonstrating that human judgments match Bayesian inference often uses analytic probabilistic models. For example, when modelling how people integrate two sources of uncertain perceptual information, researchers have used priors, likelihoods, and loss functions that result in a Bayesian solution that is simply the weighted average of the observable information (Ernst & Banks, 2002; Körding & Wolpert, 2004).

However, because of the inherent complexity of many physical processes and aspects of physical reasoning, analytic solutions cannot be used to solve many real-world problems. Instead we must consider what sorts of approximations the mind makes – how knowledge is represented, accessed and used in behavior – in order to efficiently solve these problems. Consider the simple, analytically tractable case of determining the relative mass of two rigid objects from their velocities before and after they collided with each other. While there are simple algebraic expressions for calculating the relative mass if the velocities are known, accounting for perceptual uncertainty in these situations greatly complicates the problem and makes pure analytic solutions intractable (Sanborn, Mansinghka, & Griffiths, 2013). These equations become more difficult to solve as the complexity of the system increases. For example, it's impossible to analytically predict the state of a system with three objects colliding (Diacu, 1996), much less precisely characterize systems with complex dynamics like fluids. Yet people have no problems stacking multiple dishes on top of each other, and regularly pour liquids from one container to another.

How then can people do probabilistic physical reasoning? Approximations of some sort seem mandatory. Following from the work in the chapter on Rational Process Models (Chapter 11), it is useful to look at the approximations to probabilistic inference from computer science and statistics which have been used as algorithmic models of human behavior in tasks such as categorization, decision making, and causal inference. These algorithms provide a tractable way of performing probabilistic inference, and also make systematic errors that often match the errors people make.

Perhaps the simplest rational process model for probabilistic physical reasoning is the exemplar model (Shi, Griffiths, Feldman, & Sanborn, 2010). Instead of maintaining an internal physical model of the world, probabilistic physical reasoning could instead be performed by remembering previous experiences and weighing them according to their similarity to the current situation. In simple tasks such as inferring which of a pair of colliding objects is heavier by observing their movement, a weighted average of only 50 prior experiences captured human-level performance across various settings of the underlying physical variables (Sanborn et al., 2013). But in more complex domains (e.g., predicting whether and in what direction a stack of blocks will fall) the number of possible object configurations is very large. Yet even in such domains we can still predict what will happen for configurations of objects that we have never seen before, suggesting that the exemplar model cannot explain much complex physical reasoning. As we outline below, people seem to represent the external physical world with an internal physical model that supports Bayesian inference. To make this Bayesian inference tractable, the mind might use a number of approximations, including model-based sampling, learning a recognition model for rapid inference, or using an approximate form of the physical model itself.

The ecological nature of physical reasoning

Physical reasoning is an attractive domain for studying how cognition uses complex, probabilistic generative models for three reasons. First, people have extensive experience with the physical world. Starting from infancy, we grow our understanding of physics from the building blocks of “core knowledge” (Spelke & Kinzler, 2007) to mature physical intuitions according to systematic developmental trajectories (Spelke, Breinlinger, Macomber, & Jacobson, 1992), driven by consistent changes in the way that infants interact with the world (e.g., developing motor skills to grasp objects; Baillargeon, 2002). Thus, by adulthood we would expect that interactions with the world should be guided by consistent physical intuitions that are compatible with accurate, Newtonian principles.[1]

Second, as researchers, we have access to normative computational models that can determine what the future state of a scene will be. This is in contrast to other instances of probabilistic cognition that rely on rich generative models (e.g., social cognition) for which it is challenging or impossible to determine normative accounts of how the world behaves. Access to this ground truth allows us to study when human inferences might deviate from the true future state of the world, and whether these errors might be the result of a rational inference process (e.g., Sanborn et al., 2013).

Finally, there are a set of computational models that serve as proxies for understanding how people simulate physics. At the core of any probabilistic model of cognition is the forward causal model that predicts how causes give rise to effects. This forward model allows us to calculate likelihoods and posterior distributions (see Chapter 3). If researchers want to model human

[1] While there are many instances of human physical reasoning that rely on incorrect principles (e.g., Caramazza, McCloskey, & Green, 1981; Gilden & Proffitt, 1989; McCloskey, Caramazza, & Green, 1980; Vasta & Liben, 1996), these errors may be based on a separate cognitive system that is used for more abstract problems. For further discussion, see Smith, Battaglia, and Vul (2018) and the section on “Errors in physical reasoning” later in this chapter.

physical reasoning in a probabilistic framework, they need a causal model that approximates the way the world works. Fortunately there exist a suite of models that are designed to approximate realistic physical interactions: computer physics engines, such as those in games and graphics software. Using these game engines to approximate the cognitive systems underlying physical reasoning has led to successful modeling of human physical predictions (Battaglia et al., 2013; Gerstenberg, Peterson, et al., 2017; Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2013), and the shortcuts that game engine designers have taken to model physics both realistically and quickly have provided ideas for how the mind performs efficient approximations of physics (Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

A mental model of physics

A key component of probabilistic cognition is the causal forward model that allows us to make inferences by understanding how the world works. For instance, when two objects collide we can reason about unobserved variables (the masses) based on observed variables (the trajectories) (Sanborn et al., 2013). This can be considered a simple instantiation of Bayes' rule, where we reason about the causes (c) based on the observed effects (e):

$$P(c|e) \propto P(e|c)P(c) \tag{1}$$

A crucial part of this equation is the likelihood model $P(e|c)$ which requires understanding how effects follow from causes – for example, how likely is it that we would observe the objects' trajectories for a given specification of the objects' masses? This likelihood can be instantiated by mental models of the world that provide us with information of how causes translate into effects (Craik, 1943), potentially using a mechanism of approximate probabilistic simulation. But how are these mental models for physical reasoning structured?

Extending prior research into spatial reasoning via continuous simulation, recent work has suggested a method for performing this model-based physical reasoning, namely that people have an Intuitive Physics Engine (IPE) that can simulate the world in a way similar to the game physics engines underlying many modern video games. According to this theory, the IPE takes a mental representation of the world and iteratively steps it forwards in time using approximately correct physical principles. However, while game physics engines are deterministic, the IPE is probabilistic in order to account for uncertainty in both initial world conditions and physical dynamics: people can never be perfectly certain of exactly how heavy an object is or how collisions will resolve. The IPE therefore provides us with a belief distribution over possible futures, such as where a thrown ball will end up, or the range of ways that a stack of blocks might topple. We define the IPE as Φ, which can transform a world state s at a given time into a distribution of future world states:

$$s^{t+1} \sim \Phi(s^t) \tag{2}$$

This belief distribution can be used as an input to other probabilistic cognitive models, forming a bridge between perception and other cognitive systems.

Mental simulation and spatial reasoning

Many theories suggest that spatial reasoning relies on representations that contain the same spatial information as real-world objects (Kosslyn, Ball, & Reiser, 1978, but c.f. Pylyshyn, 2002 for alternate theories on the nature of spatial representations). These spatial representations can be transformed via simulation: transforming the mental representations in a way similar to how their real-world counterparts would change through time. For instance, if we are asked to determine if two shapes are the same, the time it takes to make this judgment is related to the time it would take to rotate the shapes into alignment, suggesting we are mentally performing this rotation (Shepard & Metzler, 1971). If we are asked whether two edges of an unfolded paper cube will touch when refolded, our reaction times are related to the time it would take to fold the cube enough to check those edges (Shepard & Feng, 1972).

This mental transformation has two crucial components. First, the mental representations and transformations that underlie this simulation must reflect the objects and transformations that exist in the world (Fisher, 2006). If we wish to use simulation to understand how the world will unfold, this correspondence is necessary to ensure the results of our simulations approximate reality. Second, simulation acts in a step-wise fashion: one cannot predict a future state of the world without predicting intermediate states (Moulton & Kosslyn, 2009).

The same cognitive systems that let us mentally traverse through space or rotate objects might also include the capability of understanding how objects interact. Indeed, mental simulation underlies reasoning about mechanical events: our speed of reasoning about the kinematics of pulley systems depends on the number of components that must be set in motion (Hegarty, 1992), and the time it takes to judge how turning a gear in a chain will affect gears further in the chain depends on the number of intervening gears (until people discover rules that can shortcut this process; Schwartz & Black, 1996b). However, while these tasks do involve reasoning about physical events, they could be accomplished either by piece-wise simulation, or by sequential reasoning about the components (e.g., using a causal logic to assess the interaction between gear A and gear B, then gear B and gear C, etc.). We therefore turn to instances of physics where the continuous dynamics of the scene are important – understanding how objects collide, fluids pour, or things fly through the air – and discuss how a simulator that includes physical principles accounts well for human judgments about these scenarios.

The Intuitive Physics Engine

Motivated by previous theories of mental models underlying spatial and mechanical reasoning, Battaglia et al. (2013) proposed that human predictions about physical dynamics also utilize a simulation-based mental model, which they termed the Intuitive Physics Engine (IPE). While this mental model is theorized to reproduce the dynamics of the world well enough to make useful predictions (Sanborn et al., 2013; Smith et al., 2018), it is not supposed to perform these calculations analytically according to idealized physics; instead, the IPE is suggested to “favor speed and generality over the degree of precision needed in engineering problems” (Battaglia et al., 2013, pg. 18,328). These constraints are also found in a similar class of problems: modeling physics for video games, which require dynamics that are good enough to be acceptable to the game players, but also fast enough to run in real-time. These game physics engines function by eschewing analytic solutions, and instead simulating physics in a step-wise fashion with state transition functions that are locally consistent without explicitly modeling fundamental physical properties (e.g., conservation of energy; Gregory, 2014). The IPE is theorized to function in a similar fashion, using step-wise, approximate physical principles to model the world (Ullman et al., 2017).

Similar to game physics engines, the IPE takes as input a description of the state of the

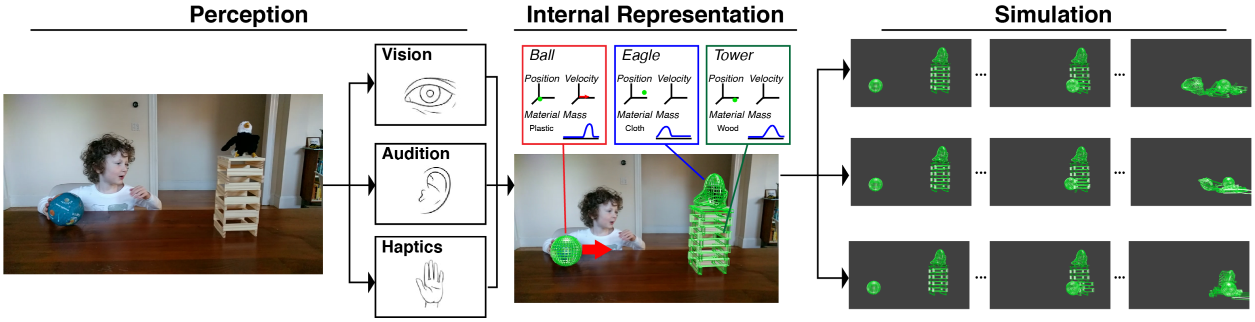

Figure 1. People perceive a scene through multiple sensory modalities (*left*) to form an internal representation of the world. This is an object-centric representation, containing probabilistic information about the locations, extents, and properties of objects (*center*). The Intuitive Physics Engine uses this representation to stochastically simulate ways the world might unfold using approximately accurate dynamics (*right*). These simulations give rise to a range of possible future states of the world that feed into other cognitive systems to make predictions, decisions, etc.

world, and yields as output simulations of hypothesized future world states. These state representations are comprised of a set of object descriptions, as objects are a basic mental building block (Spelke et al., 1992). Each object representation describes not just the shape, position, or motion of the object, but also latent properties such as mass or friction. Put together, the full state representation is similar to those used by computer-aided design programs to represent scenes, but includes additional information needed to understand the causal mechanisms that describe how the scene should unfold. However, unlike computer representations of scenes, mental representations have different memory limitations and will not include all items in a scene; instead the mind may represent only a limited set of objects that are in motion and relevant to the judgments we must make (Ullman et al., 2017).

These object representations are multi-modal, drawing on information from vision, audition, and touch. There is ample evidence that we can integrate information from vision and audition (Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003) or haptics (Ernst & Banks, 2002; Yildirim & Jacobs, 2013) to make non-physical judgments, which suggests that information from each of these modalities is integrated into a single representation in the brain (Erdogan, Chen, Garcea, Mahon, & Jacobs, 2016; Taylor, Moss, Stamatakis, & Tyler, 2006). Because the IPE relies on these integrated representations, it can also make predictions not just about how physics will transform the visual location of objects over time, but also what will be heard or felt. Conditioning on auditory information allows us to reason about material properties based on the sound of a collision (Traer & McDermott, 2016), infer the number and type of objects in an opaque box that is shaken (Siegel, Magid, Tenenbaum, & Schulz, 2014), or figure out in which hole a ball was dropped in a plinko box by integrating the sequence of sounds with information about where obstacles are positioned in the box (Gerstenberg, Siegel, & Tenenbaum, 2021).

One crucial difference between game physics engines and the IPE is that while game physics engines are deterministic, both the inputs and outputs of the IPE are belief distributions over states of the world. This distribution of beliefs over world states S comes from two sources. First, there is perceptual uncertainty in constructing mental models of the world: we are unable to exactly perceive the properties of objects given our sensory input (such as their location and velocity). In addition, the state transitions within the IPE are themselves stochastic, especially around physical events such as collisions (Smith & Vul, 2013).

Thus the IPE can be thought of as a stochastic transition function over hypothetical world states. Because both the input and the output of this model are of the same form, the same queries on the current belief state of the world can be applied to hypothetical belief states – for example, 'Where is the ball now?' is the same function applied to current beliefs as 'Where will the ball go after it is tossed?' is to predictions of future world states. Thus we can define world state queries Q such that the query on the current world state ($Q(S)$) and the query on the output of the IPE ($Q(\Phi(S))$) produce similar types of output. This provides a key link between perception and higher level cognition, providing generalized output about hypothetical futures that we can use for prediction, inference, planning, reasoning, and learning.

The physics engine in the brain. Within the brain, there are specialized neural regions dedicated to performing ecologically important tasks like recognizing faces (Kanwisher, McDermott, & Chun, 1997), or judging the mental states of others (Saxe & Kanwisher, 2003). Understanding and interacting with the physical world is another task important for our survival, so it might be expected that the brain dedicates cortical area to the IPE. Indeed, Fischer, Mikhael, Tenenbaum, and Kanwisher (2016) found that there are areas of the brain that respond preferentially to making predictions about, or just watching physical events. Furthermore, these brain regions encode information about physically relevant properties such as weight (Schwettmann, Tenenbaum, & Kanwisher, 2019) or stability (Pramod, Cohen, Tenenbaum, & Kanwisher, 2021), and are in fact the only parts of the brain from which this information can be decoded .

These “physics areas of the brain” are located in pre-motor/supplementary motor cortex and somatosensory association cortex, which is similar to brain regions that have been previously implicated in spatiotemporal prediction (Schubotz, 2007), motor action planning (Chouinard, Leonard, & Paus, 2005), and tool use (Goldenberg & Spatt, 2009). This further suggests that the IPE acts as an interface between perception and other cognitive modules that can be used to, for instance, plan our actions.

Errors in physical reasoning. To produce reasonably accurate predictions, the IPE is believed to transform mental representations of the world using principles that are approximate but generally capture how the world itself unfolds (Battaglia et al., 2013; Sanborn et al., 2013; Smith et al., 2018). This claim is distinct from a separate body of literature that finds significant errors in human reasoning about physical principles: that we display errors when reasoning about ballistic motion (Caramazza et al., 1981; Hecht & Bertamini, 2000), inappropriately believe that objects exiting curved tubes retain curvature in their motion (McCloskey et al., 1980), or fail to understand how water acts in a tipped container (Kalichman, 1988).

However, these studies that find errors in physical reasoning typically use abstract diagrams or ask for explanations of physical principles, both of which are thought to require more abstract, rule-based reasoning than more realistic, predictive tasks (Schwartz & Black, 1996a). Furthermore, tasks that rely on explicit reasoning about physical concepts activate a wider range of brain areas (Jack et al., 2013) than tasks that use more perceptual or action-oriented information (Fischer et al., 2016). Thus cases where people behave according to incorrect physical principles may be instances of reasoning with a different cognitive system than the IPE discussed here (for further discussion, see Hegarty, 2004; Smith et al., 2018; Zago & Lacquaniti, 2005).

This is not to say that the IPE always produces accurate predictions. As described later in this chapter, certain physical approximations produce biases and errors in predictions. Furthermore, it is possible that there are physical principles that are encountered rarely or have little impact on our predictions, and so are not accurately modeled by the IPE. However, for many scenarios with relatively simple shapes and dynamics that are presented in a realistic fashion, models that assume unbiased, accurate physical principles do a good job of explaining human physical reasoning.

Human physical reasoning

As a probabilistic generative model, the IPE supports many different ways of reasoning about the world. The simplest way is through prediction: running the model forwards on the current state of the world, to form a belief about how the world will turn out. But principles of probabilistic cognition suggest how the IPE can support various ways of reasoning about physics: inverting a generative model to form inferences about the world; reasoning about counterfactual models of the world to determine causality; conditioning on outcomes to plan our actions; updating models of the world in light of new evidence, and so on. In the following sections, we provide evidence for and explain how the IPE supports these various facets of cognition.

Prediction

Prediction is the simplest use of generative models of physics: running the IPE forwards and querying the simulated outcomes to make judgments about possible future states of the world. Here, probabilistic reasoning allows us to make graded predictions across a wide variety of scenarios. For example, we may predict how towers fall (Battaglia et al., 2013), balls bounce around (Deeb, Cesanek, & Domini, 2021; Gerstenberg, Peterson, et al., 2017; Smith et al., 2013; Smith & Vul, 2013, 2015) or roll down slopes (Ahuja & Sheinberg, 2019; Ceccarelli et al., 2018), objects fly under ballistic motion (Smith et al., 2018), and fluids pour (Bates, Yildirim, Tenenbaum, & Battaglia, 2019; Kubricht et al., 2016, 2017).

This is equivalent to developing a posterior belief over future world states (S^t) given the current belief over the world state (S^0) and the physics engine (Φ):

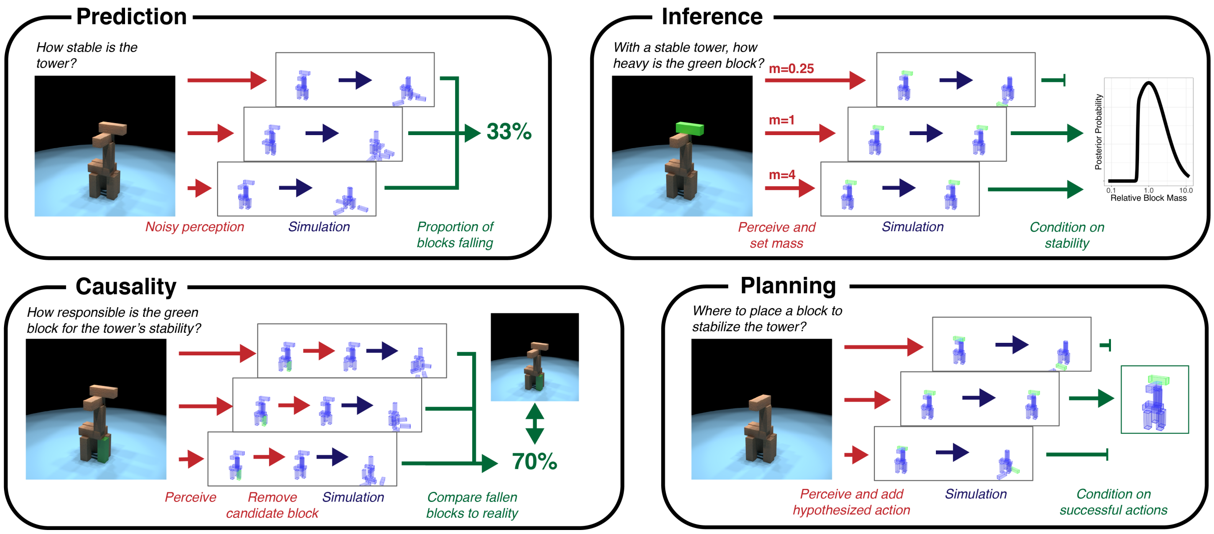
Prediction
How stable is the tower?
33%
Noisy perception
Simulation
Proportion of blocks falling
Inference
With a stable tower, how heavy is the green block?
m=0.25
m=1
m=4
Posterior Probability
0.1
1.0
10.0
Relative Block Mass
Perceive and set mass
Simulation
Condition on stability
Causality
How responsible is the green block for the tower's stability?
70%
Perceive
Remove candidate block
Simulation
Compare fallen blocks to reality
Planning
Where to place a block to stabilize the tower?
Perceive and add hypothesized action
Simulation
Condition on successful actions

Figure 2. The IPE as a generative model can support a variety of ways of reasoning about physics via probabilistic cognition. *Prediction* is running the IPE forwards and querrying the results. *Inference* requires conditioning belief based on how well a world with the relevant parameters would match observations. *Causal reasoning* requires comparing the expected result of hypothetical worlds without the causal agent to actual observations. *Planning* involves selecting actions that are expected to produce the desired outcome.

$$p(S^t) = p(S^t|S^0, \Phi)p(S^0) \tag{3}$$

Because these equations are often analytically intractable, in most cases the prior and posterior beliefs are approximated using Monte Carlo methods: treating a belief distribution as a collection of samples from a probability distribution ($S = [s_0, s_1, ...s_n]$; cf. Kahneman & Tversky, 1982). In this way each sampled state can be iteratively updated with the physics engine until a final state is reached (where Φ^* indicates iteratively applying the physics engine):

$$s_i^t = \Phi^*(s_i^0) \tag{4}$$

Battaglia et al. (2013) applied this approach to understand physical prediction. In this work, participants viewed images of block towers like those in Figure 3A, and were asked to predict whether the tower will fall or remain stable under the effects of gravity. They found that

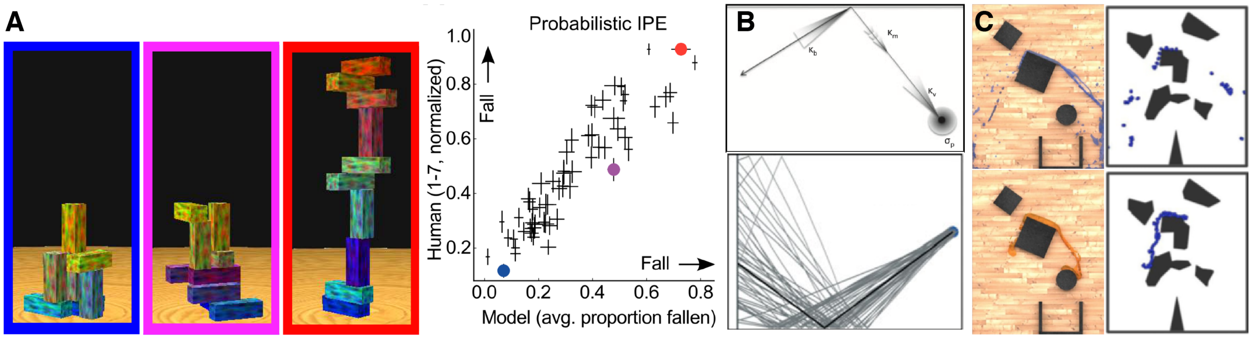

Figure 3. Instances of probabilistic physical prediction. **A.** The IPE captures predictions about general events such as stability. For instance, even though the tower with a red outline is stable, both people and an IPE model treat it as unstable (*right*) because small changes in the location or pose of almost any block will cause it to come crashing down (Battaglia et al., 2013). **B.** Uncertainty in the IPE is driven both by noise in perception and accumulating stochasticity throughout prediction (*top*), which gives rise to a distribution over possible paths that objects might take (*bottom*; Smith & Vul, 2013). **C.** Probabilistic prediction can also explain judgments of how fluids pour, by approximating the fluid with a set of interacting particles. This can differentiate between water (*top*) and honey (*bottom*) by modeling more viscous liquids as having stronger inter-particle forces (Bates et al., 2019).

participants' stability judgments could be better captured by a probabilistic simulation model than alternative, feature-based heuristics (such as the height of the tower). This model assumes that an observer has perceptual uncertainty about the exact location of the different blocks in the tower, and uses a deterministic IPE to simulate how the world will unfold under these different initial conditions. Since each initial scene will have a slightly different block configuration, the output of the IPE is a distribution over possible future scenes. Participants' judgments are then explained by aggregating the IPE's predictions across these scenes, such as the average proportion of blocks that fall. The same model also explained participants' physical intuitions across a variety of other tasks that included judging in which direction the tower will fall, or where objects would be more likely to fall off a table if it was bumped; conversely, no single feature-based heuristic could capture performance across all of these tasks.

Smith and Vul (2013) explored the extent to which noise in physical dynamics themselves affected participants' physical predictions, using a task in which participants were asked to view a ball bouncing around a computerized table and predict where that ball would travel while occluded. Like Battaglia et al. (2013), Smith and Vul assumed that participants may have perceptual uncertainty about the exact position and velocity of the ball when it disappeared behind the occluder, but they also investigated dynamic sources of uncertainty: that the ball's trajectory would be perturbed in each time step, and additionally perturbed whenever the ball collided with a wall (Fig 3B). They found that assuming uncertainty in how the physical dynamics will unfold over time was critical for explaining participants' predictions in this task, which implies that the physical transition function Φ is itself stochastic. In other experiments, this uncertainty in dynamics was also required to explain participants' judgments about their overall uncertainty about their own predictions (Smith & Vul, 2015), and how people update their predictions as a scene unfolds (Smith et al., 2013).

The proposal of the Intuitive Physics Engine has been extended beyond rigid bodies, to soft bodies, cloths, and fluids. As early as five months of age, infants demonstrate rich expectations about the dynamics of fluids and other non-solid substances, distinct from their expectations about solids (e.g. Hespos, Ferry, Anderson, Hollenbeck, & Rips, 2016). Van Assen, Barla, and Fleming (2018) found that the human visual system supports accurate inferences about fluid viscosity, which can be modeled as hierarchical estimation over mid-level visual features, such as "compactness", "elongation", "pulsing", and "clumping."

Recent work has suggested that people understand these fluid dynamics using simulation (Bates et al., 2019; Kubricht et al., 2016, 2017). Bates et al. (2019) asked participants to predict how liquids with different viscosities (water and honey) would flow down a set of obstacles, and judge what proportion of that liquid would fall into a bucket on the ground. They found that participants' predictions were well-approximated by a model that captures the complex dynamics underlying fluid motion through representing the liquid by a number of interacting particles. The results showed that participants' predictions were sensitive to the liquid's viscosity, making different predictions for how honey will flow, or how water will spill (Fig. 3C). Relatedly, a model of fluid dynamics with uncertain viscosity was used to explain people's intuitions about the angle at which a filled container would start to pour out a liquid (Kubricht et al., 2016) or sand (Kubricht et al., 2017, but see also Schwartz & Black, 1999).

Inference

It is clear how to use a forward model like the IPE to make predictions about the world: start with a set of initial conditions and run the IPE forwards. However, people can make inferences about the hidden states of a physical system just by observing how it unfolds, which requires using the IPE to make judgments in the opposite direction.

These inferences are naturally captured by Bayesian models of cognition. Here we define a set of latent properties (l) that may not be directly observable (e.g., object weight or elasticity), and observable properties (o) that may or may not change over time, such as object shape, position, or velocity. Thus a scene is a collection of latent and observed properties ($s^t = [o^t, l]$). After watching a scene unfold, posterior beliefs over the latent properties can be calculated by a simple application of Bayes' rule, conditioned on how the observed scene properties have unfolded:

$$p(l|o^t) \propto p(o^t|l, o^0, \Phi)p(o^0|l)p(l) \tag{5}$$

Sanborn et al. (2013) demonstrates how this approach can explain biases in judgment arising from human physical inferences. When observing two rigid objects colliding on a computer screen, people can infer the relative masses of the objects from observing their velocities, which requires reasoning backwards from these observed velocities to the masses that would have caused that collision. These mass judgments have been found to depend on the elasticity of the collision: when the collision is especially bouncy, people are more likely to correctly judge the heavier object to be heavier than when they observe a less elastic collision between two objects of the same masses. But according to the laws of mechanics, the relative masses of two objects can be calculated just from observations of the starting and ending velocities and should not depend on the elasticity of the collision. This dependence on an "irrelevant" variable has in the past been taken as evidence that we do not use accurate physical principles in these situations (Gilden & Proffitt, 1989; Todd & Warren Jr, 1982). However, viewing this mass judgment through the lens of probabilistic reasoning shows that the sensitivity to elasticity is not necessarily due to a simple heuristic or errors in understanding Newton's laws of

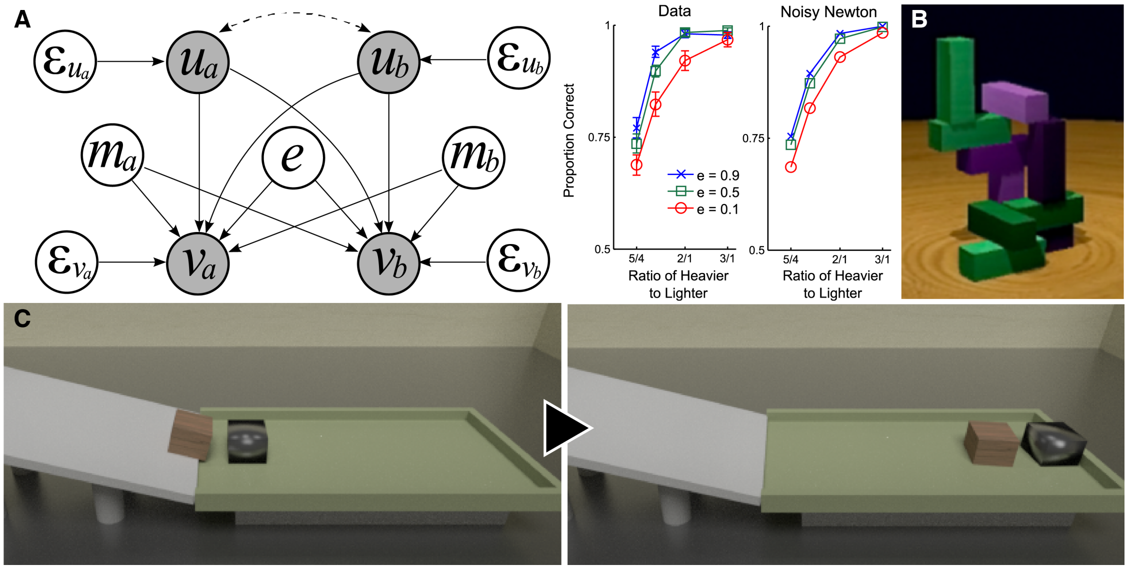

Figure 4. Probabilistic models support our ability to make judgments about latent physical properties such as mass. **A.** A graphical model used to infer the masses (m) and elasticity (e) of two blocks colliding based on the initial and final velocities $(u$ & $v)$ which are perturbed by perceptual noise (ϵ; *left*). This noisy inference model explains why people's mass judgments are biased by the elasticity of the collision (*right;* Sanborn et al., 2013). **B.** Since the tower is stable, we judge that the purple blocks must be heavier than the green blocks, because if they were not we would expect some of the blocks to fall (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). **C.** We might expect wooden blocks to be lighter than iron ones; however, if we see a block with a wood texture launching a block with an iron texture, we quickly update our beliefs about their relative weights (Yildirim, Smith, Belledonne, Wu, & Tenenbaum, 2018).

motion. Instead, collisions with slower speeds (which result from inelastic collisions) are simply harder to distinguish than collisions with faster speeds as a result of perceptual uncertainty. Similarly, people seem to be biased towards assuming that objects in motion are heavier than stationary objects (Stocker & Simoncelli, 2006); adding a prior expectation that objects move slowly results in an interaction with Newtonian mechanics that captures this bias. In this way, human judgments are consistent with Bayesian inference using an accurate model of collision dynamics (Fig. 4A; Sanborn, 2014; Sanborn et al., 2013).

People are also able to infer relative masses from scenes with more complex arrangements of objects – for example, towers of blocks (Fig. 4B) – and even update their beliefs about these relative masses across trials (Hamrick et al., 2016). These impressive feats of physical inference are not limited to vision, or to adults. Before they are one year old, infants understand that objects that compress a pillow are heavier than those that don't (Hauf, Paulus, & Baillargeon, 2012). And by shaking a box, children can infer what objects there are inside, and how many of them there are. Children can even use information about what they would *expect* to hear to determine how difficult a discrimination task would be, without having to physically shake the box. For instance, children know that two different pencils will make similar noises when shaken in a box, and so this is a difficult discrimination task, but a pencil and a cotton ball will make distinct noises and so presents an easier choice (Siegel et al., 2014).

Inferences about physical properties can in turn recalibrate the simple perceptual judgments on which they seem to be based. For example, if people see a slope with a shallow slant, but observe a ball bouncing off of the slope as if it were steep, they will adjust their perception of the orientation of the slope to be steeper, consistent with the behavior of the ball (Scarfe & Glennerster, 2014). This inference suggests that people use physical inference to build internal world representations that are consistent between their direct perception and their observations of dynamics.

Causal reasoning

Two billiard balls, ball A and ball B, collide with one another, and ball B goes into the pocket of the billiard table. Did ball A cause ball B to go into the pocket? Is it sufficient to notice that the two balls collided to answer this question about causation, or is more required? In philosophy there are two large families of theories that try to analyze what causation is. According to *process theories* of causation, causes bring about effects via a spatio-temporal contiguous process, for example, via the transmission of physical force (Dowe, 2000). According to *dependence theories* of causation, causes and effects are related via probabilistic or counterfactual dependence, such that for c to qualify as a cause of event e, e would not have happened if c hadn't happened (see Gerstenberg & Tenenbaum, 2017; Waldmann, 2017).

Both of these families of theories have had a large influence on psychological theorizing about causation. The *force dynamics model* developed by Wolff (2007) is a process theory that suggests that people judge an event to be causal based on the force transferred between the agent and patient. For example, to decide whether ball A was the cause of ball B going into the pocket, this theory suggests that we look at the configuration of forces associated with the patient and agent at the time of collision. This theory has been used to map various force configurations onto descriptions like ‘caused’ or ‘helped’ (Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010). Crucially, the force dynamics model suggests that people consider only what actually happened in order to judge whether an event was causal.

Dependence theories, on the other hand, predict that people’s judgments about causality are based on what might have happened in a counterfactual situation in which the causal event had been absent or different. The belief in what would have happened is often represented as a distribution over possible alternative outcomes, and many variants of probabilistic theories of causation exist that aim to capture people’s inferences about the strength of a relationship between putative cause and effect (Cheng, 1997; Griffiths & Tenenbaum, 2005; Jenkins & Ward, 1965). Counterfactual theories of causation naturally capture causal relationships between particular sets of events, such as whether the bump of the table caused the tower to fall, or whether the gust of wind that happened at the same time would have been sufficient to bring about the same result. These theories posit that c is a cause of event e to the extent that a counterfactual outcome e' would be different if c were removed from the scene s:

$$\text{CAUSE}(c \rightarrow e) \propto P(e' \neq e | s, \text{remove}(c)) \tag{6}$$

Gerstenberg, Goodman, Lagnado, and Tenenbaum (2021) developed the *counterfactual simulation model* of causal judgment to quantitatively capture dependence theories. According to this model, people make causal judgments by comparing what actually happened with what would have happened in a relevant counterfactual situation. For example, when asked to say whether ball A caused ball B to go into the pocket, the model not only considers that the two balls collided and that ball B went into the pocket, it also considers what would have happened if ball A hadn’t been present in the scene. The model predicts that an observer’s causal judgments will increase the more certain she is that the outcome would have been different if the cause hadn’t

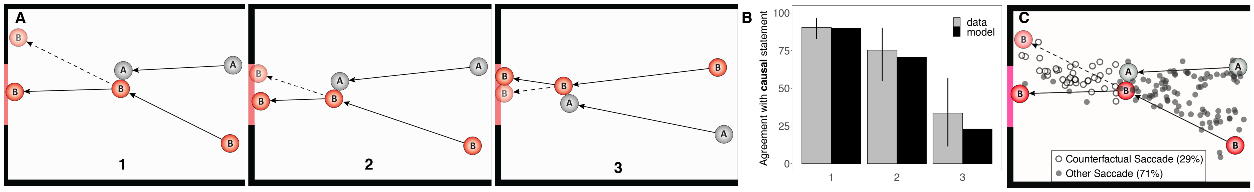

Figure 5. Probabilistic models allow us to determine what *might* have happened in the absence of a possible cause. **A.** Instances where ball A certainly (1), maybe (2), or did not (3) cause ball B to go through the gate. **B.** The counterfactual simulation model suggests that we make judgments about ball A's causal relevance by simulating what would have happened to ball B if ball A had not been there and comparing the outcome in this counterfactual situation to what actually happened. This model predicts human causal judgments well. **C.** Supporting this theory, when making causal judgments people spontaneously look towards where ball B would have gone without ball A (Gerstenberg, Peterson, et al., 2017).

been present in the scene (see Fig. 5A&B).

It's worth making explicit what's assumed to be involved in this process. In order to make causal judgments under a counterfactual theory, people first observe what actually happened. They then go back in time (mentally) and make a change to the scene in order to undo the causal event of interest (e.g. by mentally "removing" the candidate cause ball from the scene). Finally, they predict what the outcome in this counterfactual situation would have been through simulating the counterfactual course of events (see Fig. 2). This distribution over different counterfactuals arises naturally from the Intuitive Physics Engine. In some of the counterfactual scenarios, the outcome might be the same as what actually happened (i.e. ball B would still have gone into the pocket even if ball A hadn't been there; example 3 in Figure 5A), whereas in others, the outcome might have been different. People's causal judgments were well-explained by the counterfactual simulation model's uncertainty about whether the cause made a difference to whether or not the outcome happened. The more certain participants were that the outcome would have been different, the more they said that the candidate caused the outcome to happen (Gerstenberg, Goodman, et al., 2021).

Gerstenberg, Peterson, et al. (2017) tested a key prediction of the counterfactual simulation model: that people reach their causal judgment by spontaneously simulating what would have happened if the cause hadn't been present in the scene. They asked people to watch video clips of two balls colliding (A and B) and judge whether ball A caused or prevented ball B from going into a goal while using eye-tracking to determine where people looked as they made these judgments. As predicted by the model, participants looked not only at the balls, but also where ball B *would have gone* if ball A had not been in the scene. Importantly, these same eye-movements were not observed in a condition in which participants were asked only to make a judgment about what actually happened. So, the counterfactual simulations were specifically recruited in the service of making causal judgments, but without any explicit instruction in the experiment to consider counterfactual contrasts. Extensions of the counterfactual simulation model have shown how it captures people's judgments about whether something almost happened as a function of how much a causally relevant variable would have needed to be changed (e.g., the force with which a ball is kicked; Gerstenberg & Tenenbaum, 2016), and to what extent a single block in a tower is responsible for the tower's stability (by simulating what would happen if the block was removed from the tower; Gerstenberg, Zhou, Smith, & Tenenbaum, 2017).

Planning and action selection

Being able to make predictions and inferences about the physical world is about more than reasoning: it also supports rich interaction with physical systems. Specifically, a model of the physical world like the IPE can also be used to choose the best sequence of actions to take in a given scenario. This process of action selection on the basis of a model is referred to as *planning* (Sutton & Barto, 2018), and has been found to occur in the context of physical reasoning at multiple levels of abstraction, ranging from low-level motor control to high-level problem solving.

A large body of work has shown that the motor system represents forward models of how motor commands affect the motion of our bodies, the dynamics of external objects, and how our bodies might interact with those objects (Davidson & Wolpert, 2005; Flanagan & Wing, 1997; Kawato, 1999; Miall & Wolpert, 1996; Wolpert & Kawato, 1998; Wolpert, Miall, & Kawato, 1998). These forward models are used by the motor system to compute optimal actions or trajectories as follows. First, the forward models estimate possible current states in the world, either through

Bayesian inference (Wolpert, 2007) or through a filtering procedure like a Kalman filter (Grush, 2004). After an action is taken, these forward models compute expectations about what the next state of the world will be, and combine these expectations with actual sensory data to compute a posterior distribution over states. This distribution over states can be used to compute the expected cost or reward of possible actions, marginalizing over all possible states. Action selection consists of computing this expected cost (L) for all possible actions ($a \in A$), and then choosing the action (a^*) with the lowest cost (including action costs and costs of not accomplishing our goals) based on the IPE ($\Phi^*(S)$) across all plausible states ($s \in S$; Wolpert, 2007):

$$a^* = \arg\min_{a \in A} \sum_{s \in S} L(a, \Phi^*(s))p(s) \qquad (7)$$

In addition to cases where the current state of the world is uncertain, forward models also aid in computing the costs of actions when *future* states are uncertain. For example, Dasgupta, Smith, Schulz, Tenenbaum, and Gershman (2018) showed that when trying to launch a ball into a goal, people make predictions about where the ball will end up given a particular action. Where the ball ends up determines the utility of the action: if the ball makes it into the goal, there is net positive utility for accomplishing the objective, while if the ball misses the goal, a cost is incurred. To actually choose which actions to evaluate, Dasgupta et al. (2018) used a model of decision-making known as Bayesian optimization (Hernández-Lobato, Hoffman, & Ghahramani, 2014) and showed that this model not only predicted people's action evaluations, but also captured how they combined information from both mental simulations and real physical experiments. S. Li et al. (2019) also showed how physical simulations can be used to aid in computing an intrinsic reward that encourages exploratory behaviors necessary to uncovering causal properties of a physical system, similar to those produced by human participants.

However, even with a model that provides an estimate of the utility of an action, it is not always clear which actions should be considered in the first place: there are always many things we *could* do, but the vast majority of those actions will not be useful. While in theory action selection can be accomplished by exploring the space of possible actions and conditioning on those that are successful (see Fig. 2, lower-right), in reality it is impossible to consider the outcome of every possible action that could be taken. Allen, Smith, and Tenenbaum (2020) studied how people choose and use tools to accomplish goals in physical problem solving with a large space of possible actions. They find characteristics of rapid trial-and-error problem solving: people search stochastically but in a structured way at first, then exploit promising solutions to quickly solve these problems. Allen et al. (2020) propose that this rapid search requires not just a model to assess actions, but also prior expectations about what general sorts of actions are likely to be successful to avoid considering useless actions, and generalization mechanisms that take in both simulated expectations and real-world observations to update posterior beliefs about what might be useful actions to take.

In physical problem-solving tasks that require multiple steps—such as stacking blocks into a tower—a model of physical dynamics can be used to score plans depending on physical constraints. For example, Yildirim, Gerstenberg, Saeed, Toussaint, and Tenenbaum (2017) examine a block-stacking task in which a set of blocks must be assembled into a given target configuration. To model this task, they first search for a symbolic plan specifying which blocks should be stacked with which hands and in what order, and then score plans according to the physical stability of the tower in each step (along with other geometric and spatial constraints). Yildirim et al. (2017) showed that this model captures how likely human participants are to use one or two hands to solve the task, suggesting that this choice in humans may also be informed by estimates of physical stability. Yildirim et al. (2019) demonstrated how an extension of the model which takes into account physical effort and physical risk accurately captures people's intuitions about how difficult it would be to build certain block towers .

Finally, it is worth noting that planning is not necessarily limited to scenarios involving physical reasoning, and an exciting direction for future work is to combine insights from the literature on non-physical planning and learning with forward physical models like the IPE. For example, hippocampal replay and preplay during spatial navigation tasks in rats strongly resemble rollouts of a forward model (Ólafsdóttir, Barry, Saleem, Hassabis, & Spiers, 2015; Pfeiffer & Foster, 2013), and various theories have suggested that this replay occurs during a consolidation process of model-based experience into model-free action policies (Mattar & Daw, 2018; Momennejad, Otto, Daw, & Norman, 2017). Related work has explored how people trade off between model-based and model-free accounts of learning in environments with non-stationary rewards (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Dolan & Dayan, 2013; Gläscher, Daw, Dayan, & O'Doherty, 2010; Keramati, Smittenaar, Dolan, & Dayan, 2016; Kool, Cushman, & Gershman, 2016). While such work examines the use of models during learning processes, other work has explored the use of planning at decision-time, looking at how people construct and traverse trees of possible future states (Huys et al., 2012; Solway & Botvinick, 2015; van Opheusden, Galbiati, Bnaya, Li, & Ma, 2017). A number of recent advances in AI suggest other possible mechanisms for model-based planning (Hamrick, 2019), which could be integrated with models like the IPE to build process-level models of physical planning and action selection. Indeed, recent work combining model-based planning with physical models has demonstrated how to build AI systems that can reason about complex physical scenes, such as deciding how to stack blocks into a tower (e.g. Bapst et al., 2019; Fazeli et al., 2019; Janner et al., 2019). Such methods, when combined specifically with an IPE model, may also prove useful in explaining how people interact with everyday physical scenes.

Learning models of physics

We discussed earlier how an IPE can be used for inference about the dynamic variables that led to a particular observation. The logic is relatively simple to see in the case of simple hypotheses over single variables. For example, if an object is knocked with a certain force in a frictionless environment, according to an approximate simulation, the mass m of that object determines its trajectory t (giving us the likelihood $p(t|m)$). We can then use a prior (m) and standard Bayesian inference to reason about the mass given a trajectory, $p(m|t) \propto p(t|m)p(m)$. However, the trajectory is determined not just by the mass, but also by a whole physics engine (Φ) that includes background assumptions about the way that dynamic variables such as mass are affected by forces, how objects interact, how collision dynamics work, how joints constrain entities, and so on. It would be more correct to state $p(m|t, \Phi) \propto p(t|m, \Phi)p(m)$.

The physics engine encapsulates our knowledge about the world in a way that goes beyond a specific situation involving, say, a particular tower of blocks. As such, other parts of the world may be targets of inference as well, using a logic that is similar to that used for inferring the mass of a single block. If we label a particular instantiation of a physics engine as $\Phi = \phi$ (assuming now that Φ includes both m and any other possible variables), the broader inference is $p(\phi|t) \propto p(t|\phi)p(\phi)$. Of course, it is highly unlikely that we are reasoning about all aspects of the physics engine in a given situation, and a hierarchical scheme can be useful here, in which the top-most level of the hierarchy assumes only the existence of objects, properties, and dynamics laws, but without assuming the specifics (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018). As one moves down the hierarchy, specific types of properties may be hypothesized and learned (e.g., the property of elasticity), as well as specific dynamic laws (e.g. a force of attraction or repulsion). Finally, at the lowest level of the hierarchy, particular parameters can be hypothesized and learned (e.g., the specific strength of the attraction).

This general notion of hierarchical learning can provide a 'blessing of abstraction' for learning over many domains. For instance, a programmer designing a new video game will often adapt an off-the-shelf physics engine for their specific purpose rather than redesigning the full machinery from scratch. In similar fashion, when learning how to play a new game or encountering a new physical situation, humans likely assume much of their already learned (or pre-packaged) routines and variables, and learn the specific parameters and functions necessary to generate the stimuli in the new situation. Encountering two-dimensional video-games for the first time likely requires modifying the higher levels of an IPE hierarchy, but once modified many new two-dimensional video-games represent 'more of the same,' at least at an abstract level. Coming up with the notion of a global force pulling things downwards may be onerous the first time, for example, but it can then be widely applied across many situations.

Even for small and simple domains, the space of possible laws and properties can be quite large (see Ullman et al., 2018, for a simple two-dimensional world with few laws and objects and many possible 'physical theories'). And even with a useful physical prior – say in the form of a reasonable posterior over the upper levels of a hierarchical IPE representation – a new physical situation will still present a learner with a hypothesis space that is too large for exact inference over all possible physical parameter settings, dynamic laws, and relevant properties. One method for exploring this space is to posit hypotheses driven by low level features of the scene (Ullman et al., 2018), then interact with the world to explicitly test those hypotheses. Bramley, Gerstenberg, Tenenbaum, and Gureckis (2018) found that when people are asked to judge, for instance, whether and which objects repelled or attracted each other, those who were able to interact with the scene performed “experiments” that provided good evidence to discriminate between the different forces, and were more likely to learn the correct hypothesis than others who passively watched a scene – even if it was the same scene generated by one of the people who interacted with the world. This active exploration also seems to be crucial for developing an understanding of the world: infants who have not yet developed the motor skill to grasp objects but are given “sticky mittens” that allow them to pick up objects will later interact with objects in a manner as sophisticated as older infants with grasping skills (Needham, Barrett, & Peterman, 2002). Together, this suggests that active learning is a method that we use to efficiently learn about and explore the world.

Efficient physical reasoning

While the previous section demonstrates the various ways in which the IPE can be used within the framework of probabilistic cognition to explain different facets of human physical reasoning, features of physics and the IPE make it such that applying generalized probabilistic algorithms to these problems is computationally intractable. First, because there are no analytic equations to describe how physics unfolds except in the most trivial scenarios (Diacu, 1996), general probabilistic prediction requires running the IPE forward a limited number of times to *approximate* the posterior belief about the future state of the world. Second, generalized probabilistic inference algorithms require applying the likelihood function – here, the IPE – hundreds or thousands of times to produce a well-formed posterior distribution, or even more if the algorithm is initialized poorly. Yet we use the IPE to make predictions and inferences about physics in real time. In this section, we describe possible shortcuts the mind might take to more efficiently approximate probabilistic physical reasoning.

Sampling simulations

A *sample* is a random value drawn from a probability distribution. Because the IPE is a probabilistic system (Battaglia et al., 2013; Smith & Vul, 2013), every simulation from the IPE is a sample from the probability distribution over future states of the world, conditioned on current observations. The most straightforward way to use samples from the IPE is through a brute-force Monte Carlo approximation, in which a large number samples are drawn from the IPE to give a reasonable expectation of the future (see Chapter 6 for further details). For example, Battaglia et al. (2013) and Smith and Vul (2013) used large numbers of simulations (48 and 500, respectively) to form predictions and explain how many people behave in aggregate. However, using such a large number of samples from the IPE seems rather at odds with limits on an individual's working memory and attention. Do people really sample tens or hundreds of mental simulations before making a decision?

There is a priori reason to think that people may not require a large number of simulations to make a decision. Vul, Goodman, Griffiths, and Tenenbaum (2014) performed a theoretical analysis asking what an optimal decision-making agent ought to do under time constraints. Specifically, if an agent has a limited amount of time to make as many decisions as possible, how many samples should be taken per decision? The answer is a trade-off between the utility of each correct decision, the amount of time it takes to draw a sample, and the reliability of each sample. Intuitively, if it takes a lot of time to take a sample, then fewer decisions can be made, thus resulting in lower utility. However, if each sample is very noisy, then decisions are more likely to be wrong and therefore it might be advantageous to take more samples. Through a formal analyses of this trade-off, Vul et al. (2014) found that in plausible scenarios it can actually be optimal for an agent to only take a single sample to support a decision. Making a decision based on a single sample also naturally explains the classic cognitive bias of probability matching: in experiments in which people are asked to predict whether a high-probability or low-probability outcome will occur, they tend to predict the outcomes according to their probabilities, rather than always predicting the high-probability outcome as they should (Vulkan, 2000).

To determine the number of samples that people require from the IPE to support physical judgments, Hamrick, Smith, Griffiths, and Vul (2015) ran an experiment in which participants had to predict whether a ball would go through a hole. Crucially, they varied the difficulty of each trial by changing the size of the hole (i.e., either small or large) or the margin by which the ball would go through or miss the hole. On some trials the ball would go through or miss the hole with high probability according to the IPE (e.g., 90% or 10% chance of going through the hole) while on others it was very unclear whether it would go through (e.g., closer to 50% chance). Participants in this experiment took longer to make judgments when the IPE predictions were very uncertain, suggesting perhaps that they were taking more samples in these cases.

Through a model of response time based on an optimal model of decision making known as the sequential probability ratio test, Hamrick et al. (2015) showed that differences in their participants' response times could be due to a process in which samples are accumulated until a particular level of confidence is reached. Through this model, they showed that while the number of samples varied across stimuli depending on their difficulty, on average the number of samples ranged from two to four per decision. These results corroborate other more informal analyses by Battaglia et al. (2013) and Hamrick et al. (2016) suggesting that their participants relied on one to six simulations from the IPE to make decisions about towers of blocks. Thus, although each individual simulation from the IPE might be expensive, these results suggest that people can—and do—rely on only a few simulations to still achieve reasonable levels of accuracy in their judgments.

A number of questions remain regarding the computational efficiency of sampling from the IPE as well. If each sample taken from the IPE is actually a noisy physical simulation, then there are additional parameters that can be set which affect the amount of time it takes to run that simulation. For example, there is a choice of how long each simulation should be run for (e.g., how many time steps). Another simulation parameter that can be adjusted is the level of detail the simulation should be run at (e.g., the length of each time step). Similar analyses of the speed-accuracy trade-off can be performed to answer these questions, and are exciting directions for future research.

Rapid inferences

A long standing tradition in psychology has been to treat perception as inference: if we have a generative model of optics, we can condition on our retinal inputs to understand how objects are segmented in the world and where they are located (Von Helmholtz, 1867). This tradition has been carried forwards to suggest that people perceive latent physical properties (e.g., mass) from dynamic scenes by conditioning those variables based on how well their observations match what they should expect to see based on their IPE with different settings of those parameters (Hamrick et al., 2016; Sanborn et al., 2013). In practice, this inference is often carried out by 'analysis-by-synthesis' (Yuille & Kersten, 2006): setting the initial conditions of the scene (e.g., the masses and densities of objects), running the IPE forwards, then perturbing those initial conditions via a process like MCMC until the predictions of the IPE match the observations. However, this approach has been criticized for being computationally infeasible for cognition, as in the general case it requires running the generative model hundreds or thousands of times to form a good posterior estimate over those latent physical variables. This approach would clearly be at odds with the findings that people use only a handful of physical simulations in most scenarios (as described in the prior section).

If the mind is to produce these inferences as rapidly as it does, it must therefore have ways of speeding up this inference process. One method for doing so is to initialize the inference process with an intelligent guess from bottom-up features (Yuille & Kersten, 2006). Poor initializations require running the generative model to assess model parameterizations that are unlikely to explain the world; conversely, a good initialization can speed up inference by ensuring each sample from the generative model is informative. Models that implement this rapid initialization via pattern recognition (using deep networks; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015) or trained features (Ullman et al., 2018) have been found to describe human inferences better than either pattern recognition or full reasoning over the space of hypotheses.

The analysis-by-synthesis approach to inference traditionally is applied to problems with a fixed amount of information – for example, judging relative masses after observing a full video of two objects colliding (Sanborn et al., 2013; Wu et al., 2015). But physical events by their nature are dynamic, unfolding over time. Yildirim et al. (2018) demonstrate that human inferences about weight change along with the unfolding observations from the world. Furthermore, they suggest that additional approximations to the inference process are required to explain how these judgments change over time. Following the theory of Rational Process Models (Griffiths, Vul, & Sanborn, 2012, Chapter 11), they suggest that these inference dynamics can be explained by a model based on particle filters, in which belief about masses is formed as a limited set of hypotheses that are tracked and updated over time. But they also propose another possible explanation: that people might have an approximate inverse IPE that can go directly from observations to latent scene causes. This is similar to other proposals for amortized inference over generative models (Le, Baydin, & Wood, 2016; Stuhlmüller, Taylor, & Goodman, 2013), which suggest that we can use our IPE to imagine scenes that can be used to train an approximate inverse model. This inverse model will be less flexible than analysis-by-synthesis, but will also be much more efficient, and therefore might be useful to have for inference tasks we must do often or quickly (e.g., judging mass in common scenarios). Determining what approximations the mind uses for online physical inferences therefore remains an open area of research.

Physics hacks and game engine approximations

Many of the approximations that are relevant for IPEs are also relevant for general efficient inference schemes, including sampling and the heuristic use of bottom-up features. However, an IPE may also contain domain-specific *conceptual* approximations, useful for physical reasoning. Engineers that develop physics-engines for video games work under the constraint of generating ‘good enough’ simulations in real time, at everyday scales. Such engineers are not working to create a high-fidelity model of fluid dynamics, or cloud mechanics, or molecular interactions, but rather to make a splash of water look reasonable enough. In order to achieve this, engineers use principled workarounds and shortcuts to overcome limitations of time, memory, and computation. Such workarounds are useful regardless of the specific implementation language or environment of the physics engine (for general game engine concepts, see Gregory, 2014). As the human mind is under similar constraints of simulating physically-plausible objects at everyday scales with a limited computational budget, we may find a convergent conceptual evolution between the workarounds and notions used in physics engines, and those used by the IPE. Below, we focus on two examples of major short-cuts and approximations, but see Ullman et al. (2017) for more detail.

Consider first the notion of *shape* as opposed to *body* in physics-engine software. The shape of an object is what is eventually rendered on the screen, while the body of an object is what is used for actual dynamic calculations and collision-detection-and-resolution. The body is often an approximation of the shape, making use of bounding boxes and convex hulls (see Fig. 6). As a simplifying example, consider a character in a video game hurling an ornate vase at a wall. While the player may see rendered on the screen an embellished object flying towards the wall (the shape), from the point of view from a physics engine, it would be a waste of resources to exactly and accurately simulate every ridge and dip in the vase as it flies and makes contact with the wall. The complex shape of the vase is represented instead by a simple convex hull (the body), or even a box. Such a hull is much easier to store in memory, and it is easier to check when this hull overlaps with another hull or surface to trigger a collision event. Physical reality does not make such a distinction, of course, but the shape/body split is a useful conceptual scheme in a game-engine that runs on hardware with finite memory and computational power, and it may be advantageous for the mind to have such a split as well. Such an approximation may also help to explain why young infants do not use detailed shape representations to track object identity as it moves in space, even though they can distinguish them perceptually (Smith et al., 2019; Ullman et al., 2017; Xu, 2005; Xu & Carey, 1996). While game engines do not often set object bodies in a dynamic way, one can imagine the mind making different body approximations depending on the computational budget and task at hand. Figuring out when a vase will strike a surface with limited time to spare, a person may approximate the vase using only a coarse bounding box. By contrast, attempting to grasp a vase by the handle would require a more fine-grain body approximation that takes into account the ‘hole’ the handle makes in the convex hull.

Another major way that game physics engines save on memory and computation is by assigning entities to the categories *static* or *dynamic*. Static items are those that are immobile – objects like the ground or walls – whereas dynamic objects can move and be affected by forces. Crucially, static objects are not treated as large dynamic masses, but instead have undefined mass and so are unaffected by collisions and other forces. As with body and shape, this distinction between static and dynamic entities obviously does not exist in real physics. But it is an extremely useful approximation from an engineering perspective (e.g., it would be wasteful to calculate the infinitesimal effect that dropping an object on the ground has on the motion of the Earth), and so one that the IPE might make use of. Such a distinction can help explain why

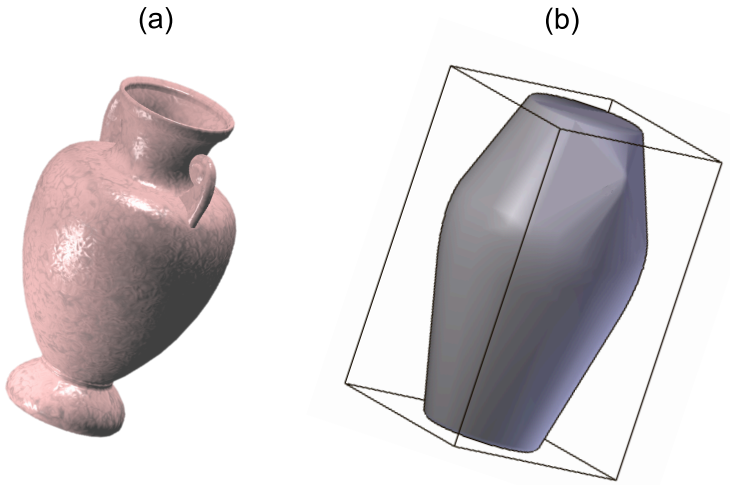

Figure 6. Difference between (a) visual shape and (b) physical body representations of an entity in a game engine. The body is an approximation of the meshes used to represent the shape, for example with a convex hull or bounding box. Images are based on an object file created by user kc8qzo on BlendSwap (https://www.blendswap.com/blend/4906).

extended surfaces are used earlier in development for navigation compared to everyday objects (Hermer & Spelke, 1994; S. A. Lee & Spelke, 2008), why shifting a wall causes changing posture and loss of balance in children and adults (D. N. Lee & Aronson, 1974), and why even very young infants expect an object made of disparate parts to move together when it is lifted, but not to take the floor with it (Spelke, Breinlinger, Jacobson, & Phillips, 1993).

There are other such concepts and shortcuts that help to organize a simulation and simplify computation, and some of them seem to explain otherwise puzzling psychological phenomena (see Ullman et al., 2017). And the inspiration can flow in the other direction – by studying the principled concepts and workarounds the IPE uses, cognitive scientists can help to develop useful tools for engineers that develop game engine simulations. Of course, it is possible that many of the concepts and workarounds in game physics engines are only the result of explicit development by engineers, with no correlate in the IPE. But given the similar need to create approximate physical representations, it is a connection worth exploring.

Future directions

Learning physical principles

There is a large body of work that has studied how infants develop an understanding of the physical world (for an overview, see Baillargeon, 1994, 2004; Kinzler & Spelke, 2007; Spelke et al., 1992, and Chapter 19). These studies find that as early as they can be tested, infants have a concept of “objects” that includes guiding principles such as the fact that solid objects do not disappear and cannot move through one another Baillargeon (1987). At the same time, many studies find that the intuitive understanding of physics develops throughout childhood. How then is the Intuitive Physics Engine learned?

The machine learning community has helped shed light on this question by introducing several approaches for training models from observed experience to predict the physical dynamics of objects and materials over time.

The “NeuroAnimator” (Grzeszczuk, Terzopoulos, & Hinton, 2000) introduced the idea of training neural networks to mimic the observed local dynamics of articulated physical systems (e.g., how a simulated robot arm’s limbs move). Two similar recent approaches—“interaction networks” (Battaglia, Pascanu, Lai, Jimenez Rezende, & Kavukcuoglu, 2016) and the “neural physics engine” (Chang, Ullman, Torralba, & Tenenbaum, 2016)—used neural networks to approximate object dynamics and force relations in physical systems which could be expressed as graphs, such as n-body gravitational systems, mass-spring systems, and rigid body dynamics with collisions. These models explicitly represented objects by the nodes of a graph, and the relations (i.e. the possibility that two objects could interact) by the edges. The models are trained by regressing from an input physical state at time t to a target physical state at time $t + 1$. The output of the model could then be fed back in as input, iteratively, to produce a long simulated trajectory.

More recent extensions of these learned physical forward models take images as input and use recurrent neural networks (Watters et al., 2017) and hierarchical representations of a physical system (Mrowca et al., 2018), and can learn to simulate or make inferences about non-rigid materials and fluids (Bouman, Xiao, Battaglia, & Freeman, 2013; Guevara et al., 2018;

Y. Li, Wu, Tedrake, Tenenbaum, & Torralba, 2018).

These models can also be used for making inferences and planning in physical systems. For example, Sanchez-Gonzalez et al. (2018) used a more powerful version of interaction networks, termed “graph networks”, to learn forward models of real and simulated robotic systems, which were then used to control the robot based on model predictions in an efficient manner. They also showed that a recurrent graph network architecture could be used to infer unobservable properties of a system from their effects on observable properties. Zheng, Luo, Wu, and Tenenbaum (2018) used a similar approach to infer properties such as mass and restitution. And Kipf, Fetaya, Wang, Welling, and Zemel (2018) introduced a probabilistic approach to inferring the structure of complex physical systems, where binary latent random variables represent the presence or absence of relations among entities.

What is perhaps most interesting is that these types of graph-based dynamics models are not specific to modeling physical systems, and can also learn non-physical dynamics, such the movements and interactions among intentional agents (Hoshen, 2017; Sukhbaatar, Szlam, & Fergus, 2016; Sun, Karlsson, Wu, Tenenbaum, & Murphy, 2019; Tacchetti et al., 2018), suggesting a method for joint physical and social prediction.

However, despite the flexibility of these learned models, they are not easily interpretable, which makes it difficult to understand how they might represent and use physical constants that are required by the IPE (e.g., gravity or mass). While there are preliminary studies of what physical knowledge might be captured by these models (e.g., Piloto et al., 2018; Riochet et al., 2018), further work is required to understand how learned models of physics capture human physical concepts.

Combining simulation with rule-based reasoning

This chapter has focused on the IPE as the forward model of physics that people use, as its stochastic nature makes it easily interpretable within the framework of probabilistic cognition. But there are also theories of physical reasoning that suggest people do not use simulation for physical reasoning, and that this reasoning is instead based on a set of axioms and logical rules (DiSessa, 1993; Hayes, 1979). These logic based theories have been used to explain how people reason about containment relationships (Davis, Marcus, & Frazier-Logue, 2017), use (biased) rules to judge whether objects will balance on a beam (Siegler, 1976), or use heuristics to predict how water will settle in a tipped container (Vasta & Liben, 1996).

Although simulation theory and rule-based reasoning make very different assumptions about the underlying representations and mental processes that support physical reasoning, they describe separate capabilities that we are able to bring to bear depending on the situation to understand the world. For instance, Smith et al. (2018) found that when people are asked to catch an object in ballistic motion, their predictions are consistent with simulations from an IPE, but when those same people are asked to draw the motion of objects in identical situations, their drawings demonstrate idiosyncratic biases. This result supports theories that suggest that we typically use simulation in scenarios that are more dynamic and realistic, but use rules and heuristics when encountering more abstract diagrams or explicit problems (see also Hegarty, 2004; Schwartz & Black, 1996a; Zago & Lacquaniti, 2005).

The cognitive systems that underlie logical reasoning are often posed as mostly deterministic, which makes them difficult to reconcile with probabilistic cognition. It is therefore important to understand how simulation and rules can be combined into a probabilistic framework. Prior work has focused on how people can learn these rules from simulation and feedback, where it is easy if the rule is physically relevant (Schwartz & Black, 1996b), but more difficult with unrelated cues (Callaway, Hamrick, & Griffiths, 2017). These findings often assume that once a good rule is learned, it will supplant the use of simulation (Schwartz & Black, 1996b).

But there are many scenarios where we do not use just simulation or just rules. For instance, there are also cases for which logical analysis of a scene provides a clear answer but people still rely in part on simulation. When predicting the motion of a ball that is contained within a box it should be easy to judge that the ball will never reach an area outside the box based on the containment relationships alone (Davis et al., 2017), but people will at least sometimes use simulation to make those judgments (Smith et al., 2013; Smith, de Peres, Vul, & Tenenbaum, 2017). Similarly, there are situations where people use rules that are biased and less accurate than physical simulation: the rules that people use for balance judgements produce biases that privilege weight over leverage for comparing torques around a center point (Siegler, 1976), but these biases cannot be derived from an IPE (Marcus & Davis, 2013). In these cases, people must choose between inaccurate but cheaper heuristics versus more accurate but more cognitively expensive simulation. While there have been initial proposals for how this trade-off is performed (e.g., based on an implicit cost / benefit comparison; Smith et al., 2018), it is an open question how people choose between and combine these different systems for physical reasoning, and how this combination of systems fits within the general framework of probabilistic cognition.

Joint physical and social reasoning

Intuitive physics and intuitive psychology deal with seemingly different domains – objects and agents, things and people. Even infants have diverging expectations when an entity is seen as a physical body compared to a perceiving agent, and some cognitive development researchers propose that different reasoning systems form two separate modules for handling these separate domains (Kinzler & Spelke, 2007), with a classification scheme that triggers different expectations depending on the type of entity that is being considered. Ongoing work in cognitive neuroscience has also identified dissociation in brain region activity when processing physical and social scenes (Fischer et al., 2016; Isik, Koldewyn, Beeler, & Kanwisher, 2017). However, even if these two domains are handled by two different computational modules, they must work in concert to produce reasonable interpretations of common scenes. Agents are physical beings that are subject to physical constraints, and these constraints help make sense of the goals, beliefs, and intentions of agents. Consider for example a simple scene in which 10-month olds see an agent jump over a barrier to get to a goal (Gergely, Nádasdy, Csibra, & Bíró, 1995). When the goal is removed, both adults and infants expect the agent to make a bee-line for the goal, rather than repeat the spatio-temporal trajectory it took previously (jumping over a now non-existent barrier). Such an expectation is obvious and intuitive, but only if we take agents to have goals, to act efficiently to achieve their goals, and – crucially for the current point – to not be able to pass through solid barriers.

Working with the framework of Bayesian Theory of Mind to intuitive psychology (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009, Chapter 14), the link between psychological and physical reasoning happens in several ways. First, physics provides the baseline *transition function* for the world, needed for planning actions. That is, in order to plan an agent needs to know $P(s'|s, a)$, the probability of moving to a new state s' conditioned on being in a specific state s and taking a specific action a (which may be not to act at all). In general such a transition function is arbitrary, and can apply to any planning context (e.g. it can describe the possible legal moves in an abstract game of tic-tac-toe), but in a real-world dynamic context this transition function is provided in part by the IPE ('If I throw this apple, what will happen?'). By inverting such a planning procedure, people can work backwards to reason about the goal that generated that plan (and see Holtzen, Zhao, Gao, Tenenbaum, & Zhu, 2016, for an implementation that infers people's hierarchical goals from videos of them moving in an everyday environment).

Second, physics provides a natural notion of cost, that can be used to estimate the reward of the agent. A great deal of psychological reasoning can be reduced to the Naive Utility Calculus (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016):

$$U(a, s) = R(s) - C(a) \tag{8}$$

Here the utility U of an agent is determined by the reward of a state s and the cost of an action a. If we have a good estimate for C, we can reason about the likely rewards that drove an agent to pay that cost (see Chapter 14, Section 1.1 for further details). There can be different types of cost, coming from mental effort, opportunity cost, temporal discounting, and so on. But a basic, natural type of cost is physical effort. The more an agent is willing to physically exert itself to get to a particular state s, the more that s must be worth. Even young infants can infer value from cost in this way, reasoning that if an agent was willing to climb a steep hill to get to Goal A, but only a shallow hill to get to Goal B, then A must be worth more to the agent than B (Liu, Ullman, Tenenbaum, & Spelke, 2017, although more work is needed to establish whether the physical effort here is related to force or distance). In a social situation, young children can use a similar calculus to reason that if a person A is unwilling to spend some small amount of physical effort to help B, then person A must not really like B (Jara-Ettinger et al., 2016). In this way, the IPE and Naive Utility Calculus can jointly provide a unified computational framework for explaining the everyday inferences we make about the plans of others given their physical constraints (see also, Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021).

Conclusion

We regularly reason about our physical world by making predictions about what will happen next, updating our beliefs about the properties of objects, and planning how we will act. While these tasks often intuitively seem effortless, performing them requires both rich generative models of the world and the capability to deal with the underlying uncertainty in perception and dynamics. Probabilistic models of cognition can help us explain how we can simulate physics under uncertainty, and how those simulations support a range of ways of reasoning about the world. Conversely, studying physical reasoning can help develop an understanding of how the mind approximates Bayesian principles in complex domains, as many of the problems we solve easily are in principle computationally intractable. Thus physical reasoning is a quintessential domain to use and extend probabilistic models of cognition.

Acknowledgements

The authors would like to thank Mario Belledonne for developing the scripts used to generate the images in Figure 2, and Matt Overlan for helpful comments and feedback.

References

Ahuja, A., & Sheinberg, D. L. (2019). Behavioral and oculomotor evidence for visual simulation of object movement. *Journal of Vision*, *19*(6), 13-13. doi: 10.1167/19.6.13

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257–262.

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. doi: 10.1073/pnas.1912341117

Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology*, *23*(5), 655-664. doi: http://dx.doi.org/10.1037/0012-1649.23.5.655

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

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

Baillargeon, R. (2004). Infants' physical world. *Current directions in psychological science*, *13*(3), 89–94.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. B. (2019). Structured agents for physical construction. *arXiv preprint arXiv:1904.03177*.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), e1007210. doi: 10.1371/journal.pcbi.1007210

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

Battaglia, P., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Josa a*, *20*(7), 1391–1397.

Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., & Kavukcuoglu, K. (2016). Interaction Networks for Learning about Objects, Relations and Physics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (p. 4502-4510). Curran Associates, Inc.

Bouman, K. L., Xiao, B., Battaglia, P., & Freeman, W. T. (2013). Estimating the material properties of fabric from video. In *Proceedings of the ieee international conference on computer vision* (pp. 1984–1991).

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive psychology*, *105*, 9–38.

Callaway, F., Hamrick, J., & Griffiths, T. (2017). *Discovering simple heuristics from mental simulation* (Preprint). Open Science Framework. doi: 10.31219/osf.io/wrqtp

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

Ceccarelli, F., La Scaleia, B., Russo, M., Cesqui, B., Gravano, S., Mezzetti, M., . . . Zago, M. (2018). Rolling Motion Along an Incline: Visual Sensitivity to the Relation Between Acceleration and Slope. *Frontiers in Neuroscience*, *12*. doi: 10.3389/fnins.2018.00406

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Chouinard, P. A., Leonard, G., & Paus, T. (2005). Role of the primary motor and dorsal premotor cortices in the anticipation of forces during object lifting. *Journal of Neuroscience*, *25*(9), 2277–2284.

Craik, K. J. W. (1943). *The nature of explanation*. Oxford, UK: University Press, Macmillan.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *bioRxiv*, 321497.

Davidson, P. R., & Wolpert, D. M. (2005). Widespread access to predictive models in the motor system: a short review. *Journal of neural engineering*, *2*(3), S313.

Davis, E., Marcus, G., & Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial Intelligence*, *248*, 46–84.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Deeb, A.-R., Cesanek, E., & Domini, F. (2021). Newtonian Predictions Are Integrated With Sensory Information in 3D Motion Perception. *Psychological Science*, *32*(2), 280–291. doi: 10.1177/0956797620966785

Diacu, F. (1996). The solution of the n-body problem. *The Mathematical Intelligencer*, *18*(3), 66–70.

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and instruction*, *10*(2-3), 105–225.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.

Erdogan, G., Chen, Q., Garcea, F. E., Mahon, B. Z., & Jacobs, R. A. (2016). Multisensory part-based representations of objects in human lateral occipital cortex. *Journal of cognitive neuroscience*, *28*(6), 869–881.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429.

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J., & Rodriguez, A. (2019). See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, *4*(26), eaav3123.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, *113*(34), E5072–E5081.

Fisher, J. C. (2006). Does simulation theory really involve simulation? *Philosophical Psychology*, *19*(4), 417–432.

Flanagan, J. R., & Wing, A. M. (1997). The role of internal models in motion planning and control: evidence from grip force adjustments during movements of hand-held loads. *Journal of Neuroscience*, *17*(4), 1519–1528.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*. doi: 10.1037/rev0000281

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731–1744.

Gerstenberg, T., Siegel, M., & Tenenbaum, J. (2021). *What happened? Reconstructing the past through vision and sound.* doi: 10.31234/osf.io/tfjdk

Gerstenberg, T., & Tenenbaum, J. B. (2016). Understanding "almost": Empirical and computational studies of near misses. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2777–2782). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 409–414). Austin, TX: Cognitive Science Society.

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

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Goldenberg, G., & Spatt, J. (2009). The neural basis of tool use. *Brain*, *132*(6), 1645–1655.

Gregory, J. (2014). *Game engine architecture*. AK Peters/CRC Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and brain sciences*, *27*(3), 377–396.

Grzeszczuk, R., Terzopoulos, D., & Hinton, G. (2000). *Neuroanimator: fast neural network emulation and control of physics-based models.* University of Toronto.

Guevara, T., Pucci, R., Taylor, N., Gutmann, M., Ramamoorthy, S., & Subr, K. (2018). To stir or not to stir: Online estimation of liquid properties for pouring actions. In *Neural information processing systems (neurips), modeling the physical world: Perception, learning, and control workshop.*

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16.

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

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. In *Cogsci.*

Hauf, P., Paulus, M., & Baillargeon, R. (2012). Infants Use Compression Information to Infer Objects' Weights: Examining Cognition, Exploration, and Prospective Action in a Preferential-Reaching Task. *Child Development*, *83*(6), 1978-1995. doi: 10.1111/j.1467-8624.2012.01824.x

Hayes, P. J. (1979). The naive physics manifesto. *Expert systems in the microelectronic age*.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 730-746. doi: 10.1037/0096-1523.26.2.730

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1084.

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

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature*, *370*(6484), 57.

Hernández-Lobato, J. M., Hoffman, M. W., & Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems* (pp. 918–926).

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

Holtzen, S., Zhao, Y., Gao, T., Tenenbaum, J. B., & Zhu, S.-C. (2016). Inferring human intent from video by sampling hierarchical plans. In *Intelligent robots and systems (iros), 2016 ieee/rsj international conference on* (pp. 1489–1496).

Hoshen, Y. (2017). Vain: Attentional multi-agent predictive modeling. In *Advances in neural information processing systems* (pp. 2701–2711).

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 201714471.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013). fmri reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385–401.

Janner, M., Levine, S., Freeman, W. T., Tenenbaum, J. B., Finn, C., & Wu, J. (2019). Reasoning about physical interactions with object-oriented prediction and planning..

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky

(Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kalichman, S. C. (1988). Individual differences in water-level task performance: A component-skills analysis. *Developmental Review*, *8*(3), 273–295.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, *17*(11), 4302–4311.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, *9*(6), 718–727.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Kinzler, K. D., & Spelke, E. S. (2007). Core systems in human cognition. *Progress in brain research*, *164*, 257–264.

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., & Zemel, R. (2018). Neural relational inference for interacting systems. *arXiv preprint arXiv:1802.04687*.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology*, *12*(8), e1005090.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of experimental psychology: Human perception and performance*, *4*(1), 47.

Kubricht, J., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1805–1810).

Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S.-C., & Lu, H. (2017). Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 700–705).

Le, T. A., Baydin, A. G., & Wood, F. (2016). Inference compilation and universal probabilistic programming. *arXiv preprint arXiv:1610.09900*.

Lee, D. N., & Aronson, E. (1974). Visual proprioceptive control of standing in human infants. *Perception & Psychophysics*, *15*(3), 529–532.

Lee, S. A., & Spelke, E. S. (2008). Children's use of geometry for reorientation. *Developmental science*, *11*(5), 743–749.

Li, S., Sun, Y., Liu, S., Wang, T., Gureckis, T., & Bramley, N. (2019). Active physical inference via reinforcement learning.

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

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(5), 1139–1141.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks*, *9*(8), 1265–1279.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2017). Offline replay supports planning: fmri evidence from reward revaluation. *bioRxiv*, 196758.

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental emulation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1521), 1273–1280.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. In *Advances in neural information processing systems* (pp. 8813–8824).

Needham, A., Barrett, T., & Peterman, K. (2002). A pick-me-up for infantsâĂŹ exploratory skills: Early simulated experiences reaching for objects using âĂŸsticky mittensâĂŹ enhances young infantsâĂŹ object exploration skills. *Infant behavior and development*, *25*(3), 279–295.

Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife*, *4*, e06063.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74.

Piloto, L., Weinstein, A., TB, D., Ahuja, A., Mirza, M., Wayne, G., … Botvinick, M. (2018). Probing Physics Knowledge Using Tools from Developmental Psychology. *arXiv:1804.01128 [cs]*.

Pramod, R., Cohen, M., Tenenbaum, J., & Kanwisher, N. (2021). *Invariant representation of physical stability in the human brain* (Preprint). Neuroscience. doi: 10.1101/2021.03.19.385641

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, *25*(02). doi: 10.1017/S0140525X02000043

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. *arXiv:1803.07616 [cs]*.

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology*, *5*, 1-7.

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

Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., & Battaglia, P. (2018). Graph networks as learnable physics engines for inference and control. *arXiv preprint arXiv:1806.01242*.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in âĂIJtheory of mindâĂİ. *Neuroimage*, *19*(4), 1835–1842.

Scarfe, P., & Glennerster, A. (2014). Humans use predictive kinematic models to calibrate visual cues to three-dimensional surface slant. *Journal of Neuroscience*, *34*(31), 10394–10401.

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in cognitive sciences*, *11*(5), 211–218.

Schwartz, D. L., & Black, J. B. (1996a). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, *30*(2), 154–219.

Schwartz, D. L., & Black, J. B. (1996b). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive science*, *20*(4), 457–497.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116–136.

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*, e46619. doi: 10.7554/eLife.46619

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive psychology*, *3*(2), 228–243.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychological Bulletin and Review*, *17*, 443-464.

Siegel, M., Magid, R., Tenenbaum, J., & Schulz, L. (2014). Black boxes: Hypothesis testing via indirect perceptual evidence. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520.

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118.

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Smith, K. A., de Peres, F. A. B., Vul, E., & Tenenbaum, J. B. (2017). Thinking inside the box:

Motion prediction in contained spaces using simulation. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3209–3214). Austin, TX: Cognitive Science Society.

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E. S., Tenenbaum, J. B., & Ullman, T. D. (2019). Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations. In *33rd Conference on Neural Information Processing Systems.* Vancouver, Canada.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2015). Prospective uncertainty: The range of possible futures in physical prediction. In *Cogsci.*

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890.

Spelke, E. S., Breinlinger, K., Jacobson, K., & Phillips, A. (1993). Gestalt relations and object perception: A developmental study. *Perception*, *22*(12), 1483–1501.

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

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585.

Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. In *Advances in neural information processing systems* (pp. 3048–3056).

Sukhbaatar, S., Szlam, A., & Fergus, R. (2016). Learning multiagent communication with backpropagation. In *Advances in neural information processing systems* (pp. 2244–2252).

Sun, C., Karlsson, P., Wu, J., Tenenbaum, J. B., & Murphy, K. (2019). Stochastic prediction of multi-agent interactions from partial observations. *arXiv preprint arXiv:1902.09641*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tacchetti, A., Song, H. F., Mediano, P. A., Zambaldi, V., Rabinowitz, N. C., Graepel, T., ... Battaglia, P. W. (2018). Relational forward models for multi-agent learning. *arXiv preprint arXiv:1809.11044*.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences*, *103*(21), 8239–8244.

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

Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, *113*(48), E7856–E7865.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, *104*, 57–82.

Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current Biology*, *28*(3), 452–458.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Cogsci.*

Vasta, R., & Liben, L. S. (1996). The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, *5*(6), 171–177.

Von Helmholtz, H. (1867). *Handbuch der physiologischen optik* (Vol. 9). Voss.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101-118.

Waldmann, M. R. (Ed.). (2017). *The oxford handbook of causal reasoning*. Oxford University Press.

Watters, N., Zoran, D., Weber, T., Battaglia, P., Pascanu, R., & Tacchetti, A. (2017). Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems* (pp. 4539–4547).

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Wolpert, D. M. (2007). Probabilistic models in human sensorimotor control. *Human movement science*, *26*(4), 511–524.

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

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in cognitive sciences*, *2*(9), 338–347.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Xu, F. (2005). Categories, kinds, and object individuation in infancy. In *Building object categories in developmental time* (pp. 81–108). Psychology Press.

Xu, F., & Carey, S. (1996). InfantsâĂŹ metaphysics: The case of numerical identity. *Cognitive psychology*, *30*(2), 111–153.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv preprint arXiv:1707.08212*.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*(2), 135–148.

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J. B., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Yildirim, I., Smith, K. A., Belledonne, M., Wu, J., & Tenenbaum, J. B. (2018). Neurocomputational modeling of human physical scene understanding. In *2018 conference on cognitive computational neuroscience.*

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). Cognitive, perceptual and action-oriented representations of falling objects. *Neuropsychologia*, *43*(2), 178–188.

Zheng, D., Luo, V., Wu, J., & Tenenbaum, J. B. (2018). Unsupervised learning of latent physical properties using perception-prediction networks. *arXiv preprint arXiv:1807.09244*.