Physics Engines as Cognitive Models of Intuitive Physical Reasoning

Ernest Davis[1] (davise@cs.nyu.edu), **Gary Marcus**[2] (gary.marcus@icloud.com)
Ethan Ludwin-Peery[3] (elp327@nyu.edu), **Todd M. Gureckis**[3] (todd.gureckis@nyu.edu)

[1]Department of Computer Science, NYU, New York [2]Robust AI

[3]Department of Psychology, NYU, New York

June 17, 2021

Abstract

Many studies have claimed that human physical reasoning consists largely of running "physics engines in the head" in which the future trajectory of the physical system under consideration is computed precisely using accurate scientific theories. In such models, uncertainty and incomplete knowledge is dealt with by sampling probabilistically over the space of possible trajectories ("Monte Carlo simulation"). We argue that such simulation-based models are too weak, in that there are many important aspects of human physical reasoning that cannot be carried out this way, or can only be carried out very inefficiently; and too strong, in that humans make large systematic errors that the models cannot account for. We conclude that simulation-based reasoning makes up at most a small part of a larger system that encompasses a wide range of additional cognitive processes.

1 Introduction

In computer science, virtually all physical reasoning is carried out using a physics engine of one kind or another. Programmers have created extremely detailed simulations[1] of the interactions of one million deformable red blood cells through capillaries (Lu, Morse, Rahimian, Stadler, & Zorin, 2019); the impact of turbulent air flow patterns on an aircraft carrier on a landing helicopter (Watson, Kelly, Owen, & White, 2019); the interaction of colliding galaxies (West, Ogden, Wallin, Sinkala, & Smith, 2020); the cosmology of the universe as a whole (DeRose et al., 2019), and the injuries to a human body caused by the explosion of an IED (Sławiński, Niezgoda, Barnat, & Wojtkowski, 2013). Software, such as NVidia PhysX, that can simulate the interactions of a range of materials, including rigid solid objects, cloth, and liquids, in real time, is publicly available and open source (Fingas, 2018). In artificial intelligence (AI) programs, simulation has been used for physical reasoning (Johnston & Williams, 2008), robotics (Kunze & Beetz, 2017; Timperley, Afzal, Katz, Hernandez, & Le Goues, 2018), motion tracking (Abella & Demircan, 2019), and planning (Zickler & Veloso, 2009).

The theory that human intuitive physical reasoning, is likewise powered by a physics engine has become popular in cognitive psychology in recent years.

For example, Battaglia et al. (2013) (p. 18327), observed that people can judge "whether a stack of dishes will topple,a branch will support a child's weight, a grocery bag is poorly packed and liable to tear or crush its contents, or a tool is firmly attached to a table or free to be lifted". To explain this ability they proposed "a model based on an 'intuitive physics engine,' a cognitive mechanism similar to computer engines that simulate rich physics in video games and graphics, but that uses approximate, probabilistic simulations to make robust and fast inferences . . . This proposal is broadly consistent with other recent proposals that intuitive physical judgments can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics."

Similarly, Ullman, Spelke, Battaglia, and Tenenbaum (2017) write, "We explore the hypothesis that many intuitive physical inferences are based on a mental physics engine that is analogous in

[1] In computer science and in much of the cognitive science literature on physical reasoning, (e.g. (Battaglia, Hamrick, & Tenenbaum, 2013), (Bates, Yildirim, Tenenbaum, & Battaglia, 2019), the word "simulation" is used to mean "a process resembling a physics engine" (as discussed below) and in this paper we use it exclusively with that meaning. Elsewhere in the cognitive psychology literature e.g (Barsalou, 2003), it is used in a more general sense

many ways to the machine physics engines used in building interactive video games. . . . We call this hypothesis the 'game engine in your head': evolution could equip infants with something like the high-level architecture used to interactively simulate the physics of virtual worlds in modern video games."

Sanborn, Mansinghka, and Griffiths (2013) proposed the strong view that "people's judgments [about physical events such as colliding objects] are based on *optimal* [emphasis added] statistical inference over a Newtonian physical model that incorporates sensory noise and intrinsic uncertainty about the physical properties of the objects being viewed . . . Combining Newtonian physics with Bayesian inference, explaining apparent deviations from precise physical law by the uncertainty in inherently ambiguous sensory data, thus seems a particularly apt way to explore the foundations of people's physical intuitions."

In this paper, we consider this view, as well as a weaker view, in which simulation as viewed a key but not unique component in physical reasoning. Hegarty (2004) for example argued that, "Mental simulations ... can be used in conjunction with non-imagery processes such as task decomposition and rule-based reasoning " Along somewhat similar lines, K. A. Smith, Dechter, Tenenbaum, and Vul (2013) argued that physical reasoning is generally carried out using simulation, but admit the possibility of exceptions:

> In some specific scenarios participants' behavior is not fit well by the simulation based model in a manner suggesting that in certain cases people may be using qualitative, rather than simulation-based, physical reasoning.

Our purpose in this paper is to discuss the scope and limits of simulation as a cognitive model of physical reasoning. To preview, we agree that something like simulation sometimes plays a role in some aspects of physical reasoning, but we also suggest that there are some severe limits on its potential scope as an explanation of human physical reasoning.

1. A "physics engine" does not denote a single, well-defined algorithm that can take any physical situation and generate accurate predictions. It is a broad and general class of algorithms. Designing an effective algorithm for a particular physical domain that meets particular specifications, in terms of accuracy and speed, often requires long and careful analysis by experts.

2. For some forms of intuitive reasoning, simulation is either extremely inefficient or entirely ineffective. There is every reason to suppose that humans use more appropriate strategies.

3. Research in this area has tended to focus on physical phenomena that are easily and naturally characterized in terms of physics engines and to ignore the many common kinds of phenomena that are not.

4. Profound systematic errors are ubiquitous in physical reasoning of all kinds. There is no reason to believe that all or most of these can be explained in terms of physics engines in the head. In particular the kinds of errors made can vary sharply between tasks, even when the underlying physical situation is identical. This suggests that they are not drawing on a single cognitive mechanism that corresponds to a correct physical theory, but rather are using a variety of task-dependent mechanisms.

5. In some forms of physical reasoning, simulation-based theories would require that naïve subjects enjoy a level of tacit scientific knowledge of a power, richness, and sophistication that is altogether implausible.

2 Simulation-based physical reasoning

We begin by reviewing some of the studies that have supported simulation-based theories of intuitive physical reasoning. Roughly speaking, these have fallen into two classes. Earlier work, on *depictive models*, focused on the way that humans consciously visualize the evolution of a physical system. More recent work has emphasized showing that subjects' answers can be modeled by a physics engine, which often generates direct judgments with no conscious visualization.

2.1 Depictive Models

One strand, published mainly over the 1990s and prior to 2005, couched important aspects of physical reasoning in terms of visualizations of behavior evolving over time. Figure 1 illustrates three typical examples from the earlier literature:

Hegarty (1992) studied how people infer the kinematics of simple pulley systems (3 pulleys, 1 or 2 ropes, 1 weight) from diagrams showing a starting state. Her primary data was eye fixation studies (i.e. the sequence in which subjects look at different parts of the diagram or the instructions), though accuracy and reaction times were also reported. In this paper Hegarty proposes a weak theory of simulation, in which subjects simulate each pairwise interaction between components of the system and then trace a chain of causality across the system, rather than attempting to visualize the workings of the system as a whole.

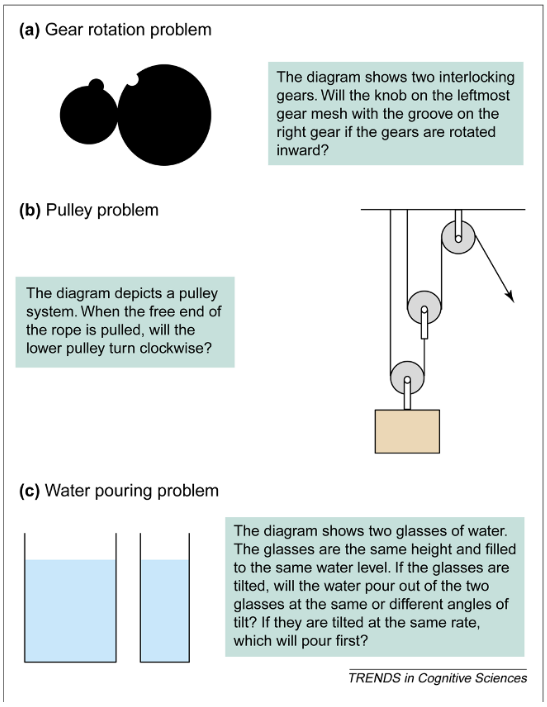
In Schwartz and Black (1996) subjects solved problems involving the motion of two gears. Two adjacent gears of different sizes are shown with either a line marked on both or a knob on one and a matching groove on the other. The subjects were asked whether, if the gears were rotated in a specified direction, the lines would come to match, or the knob would meet the groove. In all experiments, both accuracy and latency were measured, but most of the conclusions were based on latency. The primary purpose of the set of experiments as a whole was to compare the use of two possible reasoning strategies; on the one hand, visualizing the two gears, rotating synchronously; on the other hand, comparing the arc length on the rim of the two gears between the contact point in the starting state and the lines/knob/groove. (The interlocking of the gears enforces the condition that equal arc lengths of the rim go past the contact point; hence, the two markings will meet just if the arc lengths are equal.) It was conjectured that the first strategy would require a cognitive processing time that increased linearly with the required angle of rotation, but that the second strategy would be largely invariant of the angle of rotation; and the experimental data largely supported that conjecture. Various manipulations were attempted to try to get the subjects to use one strategy or another. In one experiment they were specifically instructed to use a particular strategy. In another, they were presented alternately with a realistic drawing or with a schematic resembling a geometry problem; the former encouraged the use of visualization, the latter encouraged the comparison of arc length.

Similarly, Schwartz (1999) studied the behavior of subjects who are trying to solve the following problem. "Suppose there are two glasses of the same height, one narrow and one wide, which are filled with water to equal heights. Which glass has to be tilted to a greater angle to make it pour?" Schwartz reports that subjects who answer the question without visualizing the two glasses almost always get the answer wrong (19 out of 20 subjects). However, if they visualize tilting the glasses, they are much more successful. Schwartz further tried a number of manipulations to determine when the subjects use a kinematic model and when they use a dynamic model; frankly, the relation between this distinction and the structure of his experiments is not always clear to us. The data that he used was the subjects' answers.

2.2 Newtonian physics engines

More recent work has been couched in terms of a "game engine in the head" approach (Ullman et al., 2017). A "game engine" is a computational process analogous to the physics engines used in scientific computations, computer graphics, and computer games. Broadly speaking, the physical theory that is incorporated in the engine is expressed in the form of update rule, that allows the engine to compute the complete state of the world at time $T + \Delta$ given a complete specification of its state at time T, where Δ is a small time increment. (We will discuss the significance of "complete" below.) In any particular situation, the input to the engine is the complete state of the world at time T=0; it then uses the update rule to compute the state of the world at time Δ from its state at time 0; to compute the state of the world at time 2Δ from its state at time Δ and so on. Thus, it computes an entire trajectory.

Many variants of this general idea are possible. There may be exogenous events that occur over time, such as the actions of an player in a game; in that case, the update function may have to take these into account. It may be possible to extrapolate the trajectory of the world to the next "interesting" event rather than using a fixed time increment Δ; for instance, in the bouncing ball experiment of (K. A. Smith, Dechter, et al., 2013) described below, the engine might well goes from one bounce to the next without calculating intermediate states, except to check whether the path crosses one of the target regions. Battaglia et al. (2013) suggest that the internal engine is



(a) Gear rotation problem

The diagram shows two interlocking gears. Will the knob on the leftmost gear mesh with the groove on the right gear if the gears are rotated inward?

(b) Pulley problem

The diagram depicts a pulley system. When the free end of the rope is pulled, will the lower pulley turn clockwise?

(c) Water pouring problem

The diagram shows two glasses of water. The glasses are the same height and filled to the same water level. If the glasses are tilted, will the water pour out of the two glasses at the same or different angles of tilt? If they are tilted at the same rate, which will pour first?

TRENDS in Cognitive Sciences

Figure 1: Experiments, from Hegarty (2004)

"approximate: In its mechanics rules and representations of objects, forces, and probabilities, it trades precision and veridicality for speed, generality, and the ability to make predictions that are good enough for the purposes of everyday abilities". However, the inference engine that they actually use in their model is not approximate in this sense.

Most important, if either the starting state or the update rule is partially specified but can be viewed as following a probabilistic distribution, then one can generate a random trajectory corresponding to that distribution by random sampling; this is probabilistic or Monte Carlo simulation. This partial specification of the input situation is usually attributed to limits on the precision of perception or to partial information of some other kind about the starting situation. Probabilistic models of this kind are known as "noisy Newton" models, since they follow Newtonian mechanics (or whatever exact scientific theory applies) with the addition of some random noise.

In this more recent line of work, subjects typically view and interact with a computer simulation rendered on a standard computer monitor. In some experiments (see Figure 2), the monitor displays a two-dimensional rendering of a three-dimensional situation; in other experiments, the simulation in question is itself is limited to two dimensions. Some experiments show a static picture, others show a short video clip. (In the depictive work in earlier decades, participants were generally shown static pictures printed on paper. Experiments in physical reasoning where subjects interact with or view live, 3-dimensional physical situations seem to be mostly restricted to studies with children; Won, Gross, and Firestone (2021) is an exception.)

To take one example (left panel of Figure 2), Battaglia et al. (2013) carried out one experiment in which participants were shown a tower of blocks. Participants were asked to predict whether a tower was stable and, if not, in which direction it would fall. In other experiments (right panel of Figure 2) subjects were shown a table with a number of towers of red and green blocks in various positions. Participants were told that the table would be struck at a specified point, and they were asked whether more red or more green blocks would fall off. Responses were consistent with a "noisy Newton model" in which a subject applies the Newtonian theory of rigid solid objects, and carries out probabilistic simulation, where the probabilistic element corresponds to uncertainty in the positions of the blocks.

In another study in this strand, K. A. Smith and Vul (2013) carried out an experiment in which participants catch a bouncing ball with a paddle. The trajectory of the ball is initially visible; then one side of the screen becomes occluded, and the subjects must move the paddle to a position where it will catch the ball after it has bounced around (Figure 3). The data analyzed was the relation of

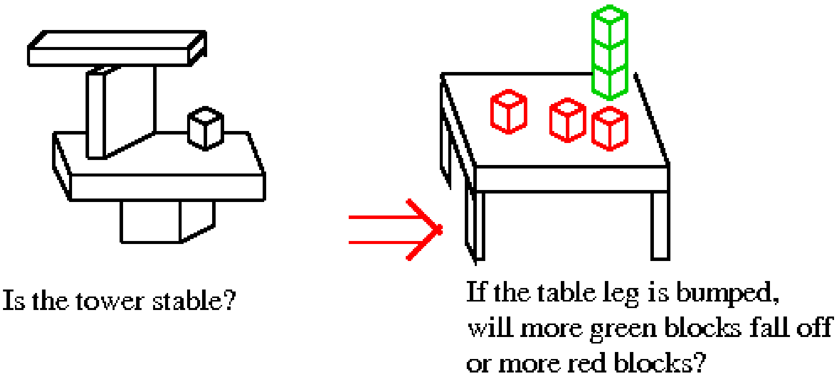

Is the tower stable?
If the table leg is bumped, will more green blocks fall off or more red blocks?

Figure 2: Experiments in Battaglia et al. (2013)

the subjects' placement of the paddle to the actual trajectory of the ball. They were able to match the data to a "noisy Newton" model in which both the perception of the ball's position and velocity and the calculation of the result of a bounce were somewhat noisy. They additionally posited a "center bias", with no apparent *a priori* justification, which they attributed to the subjects' prior expectations about the position of the ball.

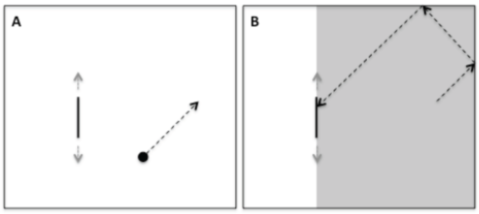


Figure 3: Diagram of a trial. (A) The ball moves unoccluded in a straight line. (B) Once the field is occluded, the ball continues to move until caught or it passes the paddle plane. From K. A. Smith and Vul (2013)

Experiments in K. A. Smith, Dechter, et al. (2013) likewise tested people's ability to predict the trajectory of a bouncing on a table with obstacles A green and a red target are marked on the table, and subjects are asked which of the two targets the ball will reach first. As they watched the simulated ball bounce, subjects were able to continuously make their best guess as to the answer, and to change their prediction when necessary. The data considered was thus the time sequence of guesses. They found that in most cases subjects' answers fit well to a noisy Newton model similar to that of K. A. Smith and Vul (2013) (they additionally posited a rather complex decision model of how subjects chose their answer.) The exceptions were cases where the correct prediction could be made purely on the basis of topological constraints; i.e. cases where any possible motion would necessarily reach one region before the other. In those cases, subjects were able to find the correct answer much earlier than their model predicted (Figure 4).

K. Smith, Battaglia, and Vul (2018) carried out experiments testing how well subjects could predict the trajectory of a pendulum bob if the pendulum is cut while swinging. Replicating the negative results of Caramazza, McCloskey, and Green (1981), they found that when subjects were asked to draw the expected trajectory, they did extremely poorly, often generating pictures that were not even qualitatively correct. However, if subjects were asked to place a bucket to catch the bob or to place a blade, they were much more accurate (Figure 5).

Noisy Newton simulation has also been used as models of human behavior in a variety of tasks involving colliding balls, including Sanborn et al. (2013), Gerstenberg and Goodman (2012), Gerstenberg, Goodman, Lagnado, and Tenenbaum (2014), Sanborn (2014), K. A. Smith and Vul (2014) and Ullman, Stuhlmüller, Goodman, and Tenenbaum (2018), study human performance on a variety of tasks involving colliding balls, including prediction, retrodiction, judgment of comparative mass, causality, counterfactuals, and learning; in predicting liquid flow (Bates et al., 2019) and (Kubricht et al., 2016); and in developmental studies (Lin et al., 2021; Téglás et al., 2011).

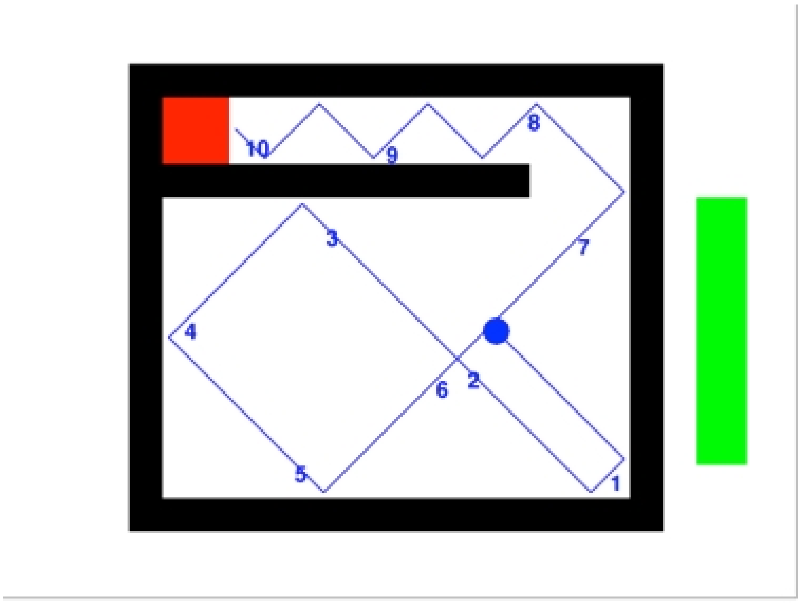


Figure 4: Simulating a bouncing ball: An example that can be solved using qualitative reasoning. From (K. A. Smith, Dechter, et al., 2013)

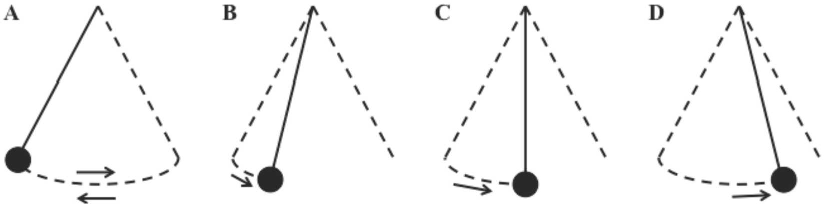


Figure 5: The four pendulums in the diagram task. Participants were asked to draw the expected path of the ball if the pendulum string were cut at each of the four points. Figure 2 of K. A. Smith, Battaglia, and Vul (2013)

3 Inherent limitations of simulation

All of this work assumes, importantly, that "a physics engine in the head" would have some kind of single, clearly-defined referent. But the literature makes no commitment to any specific physics engine (such, as e.g., the one in the widely-used Unity game engine) nor much reference to an important yet bitter reality: simulation, itself, is a complex and challenging field. There is, in computer science, no single physics engine, and no magical one-size-fits-all approach. Developing an simulation engine requires careful consideration of the physical domain, the class of problems that it will be run on, (e.g., how many objects over how long a time period), the characteristics of the input (e.g., what kind of data will be available) and the requirements on the answer (e.g., how reliable must it be, what aspects of the physical situation must it predict accurately, and what degree of accuracy is required). In order to evaluate the idea of a physics engine in the head, it is important to understand the challenges that are involved in building physics engines, and what consequently would be needed to realize such a thing in human brains.

It is easy for a non-professional to overestimate the state of the art of physical simulation, and assume that there is a plug-and-play physics engine that works for pretty much any physical situation. If computer scientists actually knew how to build a comprehensive physics engine, the idea that humans might have an instantiation of such a system might seem more plausible. But the reality is that physics engines of the sort that exist in 2021 are brittle things, when it comes to anticipating the full complexities of the real-world. Plug-and-play engines capture only narrowly-defined environments; more sophisticated applications require hard work from experts. A few seconds of realistic CGI in a disaster film may well require several person-days of work; an accurate and complex scientific computation may require several person-months. Plug-and-play physics engines are also subject to bugs and anomalies and may require careful, post hoc parameter setting to work correctly. Comparative studies of state-of-the-art physics engines have shown that each of them has areas of weakness where they give unreliable results and that none is consistently superior over all applications (Boeing & Bräunl, 2007; Chung & Pollard, 2016).

The robotics industry has invested enormous effort in building reliable simulators for robots, so that features can be tested cheaply in a virtual environment before being deployed in expensive physical robots. Nonetheless the simulated behavior is often very far from real-world behavior; roboticists call this "the reality gap" (Collins, Howard, & Leitner, 2019; Mouret & Chatzilygeroudis, 2017). Similarly, a large and important fraction of intuitive physical reasoning in people has to do with the effect of their own physical actions on the world; but the human body is very difficult to model accurately. (The integration of realistic character avatars with physics engines is a challenging problem; see, for instance, the discussion in Wang, Guo, Shugrina, and Fidler (2020).) Reasoning about the human body has received little or no attention in the literature on intuitive physics.

Even within automated reasoning of the sort found in artificial intelligence, there at least a dozen serious challenges to using simulation as a full-scale solution to physical reasoning (Davis & Marcus, 2016). Many of these challenges lead to analogous difficulties for any pure-simulation account of human physical reasoning as well. We discuss the six most important of these challenges below.

3.1 Finding an appropriate modeling approach

For a programmer building a simulation, the first hurdle in implementing a simulator is developing a domain model. In some cases, such as pendulums and blocks, this is well understood. However, finding an appropriate model can often be difficult, even for familiar objects, materials, and physical processes. Consider, for instance, cutting materials with tools. An ordinary household has perhaps a dozen kinds of tools for cutting: a few kinds of kitchen knives; more specialized kitchen equipment such as graters and peelers; a few kinds of scissors; a drill, a saw, a lawn mower, and so on. (A specialist, such as a carpenter or a surgeon, has many more.) Most people understand how they should be used and what would happen if you used the wrong tool for the material; if, for example, you tried to cut firewood with a pair of scissors. But it would be hard to find good models for these in the physics or engineering literature. Even physically simpler systems can be difficult to model plausibly; for example, doors are so difficult to get right in video game physics engines that most game designers simply make them magical (Farokhmanesh, 2021). Where the best minds in computer science and engineering haven't yet constructed detailed simulations of these sorts of situations, it might be unrealistic to presume that brains are solving routinely solving a vast array of such problems via simulation alone.

3.2 Choosing an idealization

Virtually all actually-implemented simulations represent idealizations; in some, friction is ignored; in others, three dimensions are abstracted as two. In most situations, many different idealizations are possible; and the idealization should be chosen so that, on the one hand, the calculation is not unnecessarily difficult, and on the other, the important features of the situation are preserved. Consider, for instance, the simulation of a pendulum on a string. As we have discussed, K. A. Smith, Battaglia, and Vul (2013) generated simulations in which the bob first swings on the string and then, after the string is cut, flies freely through the air. In all likelihood, the bob was modeled throughout as a point object, moving under the influence of gravity, and the string was modeled as a constraint that requires the bob to move on a circular path. The cutting of the string was modeled as an instantaneous elimination of this constraint, with the assumption that the instantaneous velocity of the bob is unchanged.

Other scenarios for a bob on a string are more complex. A bob may swing in a horizontal circle; spin on the axis of the string; rotate about its center of mass like a yo-yo, or fly through the air (Figure 6). The string itself may be taut, loose, tangled, knotted, or twisted; it may get in the way of the bob; it may even unravel or snap. Although these behaviors are familiar to anyone who has spent time playing with objects on strings, few if any existing physics engines support any but the taut and loose conditions of the string, and perhaps snapping.

Efficient reasoning about these different possible behaviors of the string and the bob requires using a variety of different idealizations. Space can be two-dimensional or three-dimensional. A bob can be idealized as a point object, a rigid object, or an elastic object. A string of length L can be idealized as an abstract constraint restricting the motion of the bob; a one-dimensional curve of length L, with or without mass; or a three-dimensional flexible object, either uniform or with some internal structure (e.g., twisted out of threads or a chain of links). Cutting the string may be instantaneous and involve no dissipation of energy; in a more realistic model, it will require finite time and involve a small dissipation of energy. Influence on the system can be limited to gravity, or can include friction and air resistance. In looking at any one simulation that has been well-worked out, it is easy to lose sight of how much careful work goes on in choosing the right idealization; as yet there is no algorithmic way to guarantee an efficient solution for arbitrary problems. Using the most realistic model possible is no panacea; highly realistic models both require more laborious calculation and more detailed information.

Useful idealizations can, indeed, be very far from the underlying physical reality. For example Téglás et al. (2011) model balls bouncing inside a container in terms of random walks; i.e. each ball moves on a random path, uninfluenced by the other balls. This is indeed a common idealization in statistical mechanics (where it is much more nearly correct, due to the enormous number of particles) but not at all a reasonable model of the dynamics of an individual ball. The motions of an actual ball in this situation are due to collisions with other balls, not due to an autonomous random process. If you are solving a billiard-balls problem, you definitely do not want to use this model.

A simulation-based model of human reasoning therefore cannot rely on simply running a uniquely-defined model for each type of object or substance. Rather, objects come with a collection of models; and choosing the right model for each object in a given situation is itself a complex reasoning task.

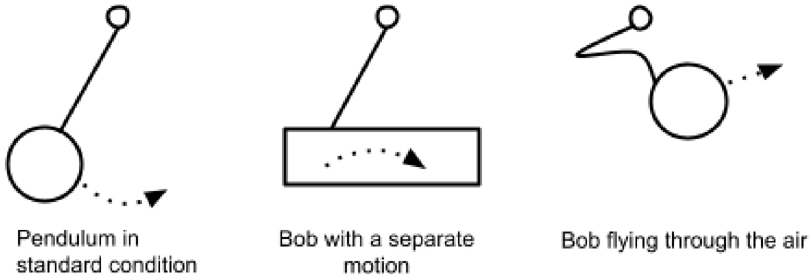


Figure 6: A pendulum in various conditions

One might, perhaps, propose a theory in which the mind has multiple physics engines, corresponding to different idealizations, used in different circumstances as appropriate. However, this is a significantly different, and much more complex, theory than the idea of a simple one-size-fits-all physics engine. The cognitive mechanism now involved in physical reasoning would now involve choosing between idealizations and, probably, combining different idealizations for different parts of a problem, which are complex and sophisticated reasoning processes.

3.3 Rapid, approximate inferences

The output of a computer simulation is invariably precise, but not always accurate. Human physical reasoning seems to have a different character, speedy, yet less precise.

Correspondingly, human reasoners often seem to have little need for the level of precision that a simulation provides. If you ride a bicycle on a bumpy road while carrying a half-full closed water canteen, all that matters is that the water stays inside the canteen, not the trajectory of the water splashing inside the canteen. Humans often seem to be able to short-cut these situations, drawing rapid inferences.

Although the exact workings of human physical reasoning remain unknown, there are many kinds of rules that in principle could allow quick inference or quick transference of results from a known situation to a new one.

- **Invariance under time and space:** If a mechanism worked in a particular way at home on Monday, it will work in the same way at work on Tuesday.
- **Invariance under irrelevant changes:** If a jar fits on a shelf, and you fill it with pennies, it will still fit on the shelf.
- **Invariance under changes of scale** (for certain physical theories, such as kinematics): A large pair of gears works in the same way as a much smaller scale model.
- **Approximation:** If a jug holds a gallon of water, then another jug of similar dimensions and shape will hold about a gallon.
- **Ordering on a relevant dimension:** If a toy fits in a box, then it will fit in a larger box, under a proper definition of "larger" (Davis, 2013).
- **Decomposition**: If a system consists of two uncoupled subsystems, then one can reason about each separately.

To take another example: suppose a subject is shown a picture of one of Hegarty's pulley problem, as in Figure 1 above, and asked whether, if the weight is made heavier, if more or less force would be required to lift it through the pulley system. One possible approach would be to calculate the force exactly for a number of different weights and compare them. But a simpler approach, and we conjecture the approach that most people would use, is to use a general rule: in any mechanical system for lifting an object, the force that must be applied is proportional to the weight to be lifted.

There can also be rules of thumb: useful if imperfect generalizations for common cases. For instance, if you spill a cup of coffee in your office, it should not be necessary to resort to simulation to infer that the coffee will not end up in some other office. Rather, one can use a rule of thumb that a small amount of liquid dropped from a moderate height onto a horizontal surface will end up not very far from the point where it was dropped, where "small amount", "moderate height", and "not very far" have some very approximate quantitative measurements.

These alternative forms of inference may do well under circumstances where simulation scales badly. Consider the following scenario: you put a number of objects in a box, close the lid, and shake it up and down violently. We now wish to infer that the objects are still in the box. Simulating the motion of the objects will become rapidly more complicated as the number of objects increases, the shapes of the objects become more complex, and the shaking of the box becomes more complex and violent. By contrast a single simple rule, "An object in a closed container remains in the container", suffices to carry out the inference. Our best guess is that such rules are an important part of the fabric of human physical reasoning.

3.4 Extra-physical information

In some cases, reasoning about a physical system can be carried out more reliably and effectively using extra-physical information. Suppose that you see a baseball pitcher throw a ball. If you simulate the motion of the ball, using the best information provided by your visual perception about the angle, velocity, and so on, of the ball when it left his hand, and factor in the imprecision of this information, you will predict that it has a rather low probability of ending up anywhere close to the batter. You would obtain better results — predicting that the ball will end up close to the strike zone, just inside it or just outside — by instead relying on the known accuracy of the pitcher, plus quite specific information about the state of play and the pitcher's willingness to risk an out-of-strike-zone pitch rather than a hit. Pure physical simulation, constrained by real-world measurement error, might produce far less accurate results.

3.5 Incomplete information

Carrying out a physical simulation is generally only possible if the geometric and physical characteristics of the initial condition are known precisely. In many common real-world situations, humans often perform physical reasoning on the basis of partial, sometimes extremely limited, information.

Perception, for example, may be imperfect or incomplete. For instance, an object may be partially occluded. (An opaque object always self-occludes its own far side from the viewer.) Knowledge of the physical situation may come from natural language text or sketches. Knowledge of aspects of the situation may come from inference; for example, if you see someone attempt unsuccessfully to pick up a suitcase, you can infer that the suitcase is unusually heavy; you can then use that inference for future prediction. Or the precise details may not have been determined yet. For instance, suppose that you are going to the furniture store to buy a dining room table. You can reason that you will not be able to carry it home on foot or riding a bicycle, even though you have not yet chosen a particular table.

Of course, no representation is truly complete or entirely precise; in any representation, some aspects are omitted, some are simplified, and some are approximated. However, the simulation algorithm requires that the initial conditions of the scenario be fully specified relative to a given level of description. That is, the representational framework specifies some number of critical relations between entities and properties of entities. A complete representation of a situation relative to that framework enumerates all the entities that are relevant to the situation, and specifies all the relations in the framework that hold between those entities. The description must be detailed and precise enough that the situation at the next time step is likewise fully specified, in the same sense.

In many cases people are able to carry out physical reasoning on the basis of information that is radically incomplete. For example, suppose that you have an eel inside a closed fish tank and you wish to infer that it remains in the tank. If we are to solve this problem by probabilistic simulation, we would need, first to understand how an eel swims, and second to simulate all kinds of possible motions of the eel and confirm that they all end with the eel inside the tank. If we do not know the mechanisms of swimming in eels, then the only way to use probabilistic simulation is to generate some random distribution of mechanisms that eels might use. Clearly, this is not psychologically plausible.

Another, similar example: You pack a shirt in a suitcase, you lock the suitcase, you check it onto a flight to Chicago. The suitcase is lost. Three days later, it turns up at the Dallas airport. It's pretty banged up, but intact. Who knows what they did to it or how it got there. However, if it's still locked, you can be sure that the shirt is still inside. In this case, it is impossible to effectively simulate because there is no information about the exogenous events involved.

Fields like forensics, archaeology, and paleontology largely depend on this kinds of inference from partial information. You find the skeleton of a prehistoric creature, and you infer that at some point it had a broken leg that healed. What else happened to the creature during its lifetime and what has happened to its skeleton since it died may be entirely unknown.

3.6 Tasks other than prediction

Simulations are most readily used for the task of generating predictions, such as where a pendulum will be at a given instant. There are, however, many other important kinds of physical reasoning tasks, to which simulation-based techniques are in general less well suited. These include: interpolation between two states at two different times; planning; inferring the shape of an object; inferring the physical properties of an object; design; and comparative analysis. All of these kinds of tasks arise constantly for people in ordinary situations; a cognitive theory of physical reasoning must therefore account for all of them. A system that could only capture prediction would miss much of the richness of human physical reasoning.

In the scientific computing community it is well known and accepted that "simulators are poorly suited for statistical inference" (Cranmer, Brehmer, & Louppe, 2020). A number of algorithmic techniques have been developed for the so-called "inverse problem" of using simulators for the specific task of estimating parameters from data, but the problem remains a very difficult one. To what extent these techniques can be adapted to the varied kinds of reasoning faced by an intelligent actor is unknown.

4 Further limitations of simulation in cognitive modeling

Even where simulation could be used in principle, there is often reason to doubt that is used in practice.

4.1 Systematic errors in physical reasoning

> *"How should I answer these questions — according to what you taught me, or how I usually think about these things?"*

– Harvard physics student, confronted with David Hestenes' test of basic physics concepts (Lambert, 2012)

There is an enormous literature demonstrating that both naïve and educated subjects make systematic, large errors in simple physical reasoning tasks. We have already discussed some of these. Tasks in which well-documented errors commonly occur include:

1. Predicting the trajectory of an object that has been moving in a circular path and is now released (McCloskey, 1983)

2. Drawing the predicted trajectory of bob on a pendulum that is cut (K. A. Smith & Vul, 2013)

3. Predicting the behavior of balance beams (Siegler, 1976)

4. Generating explanations for the workings of familiar mechanisms, such as bicycle gears (Keil, 2003)

5. Judging the relation between the masses of two colliding objects based on seeing a collision between them (Gilden & Proffitt, 1994)

6. Predicting whether two objects will in fact collide (Levillain & Bonatti, 2011)

7. Predicting whether another person can see themselves in a mirror (Lawson, 2012)

8. Predicting the behavior of a wheel (Proffitt, Kaiser, & Whelan, 1990)

The list could be extended at very great length. These kinds of errors are inherently challenging for theories of simulation based on correct physical theories. In a few cases such as (5) above, it is possible to explain the errors in terms of a "noisy Newton" theory as the consequence of taking perceptual noise into account (Sanborn, 2014). In most other cases, no such explanation has been offered and in many no such explanation seems to be possible.

Although some of these might be individually explained away, as errors of misinterpretation, in our view such errors are simply too ubiquitous to be explained away. And although "Bayesian" approaches to cognition often presume optimality as a kind of default, there is no particular reason that human cognitive processes, developed by evolutionary processes, many over only a comparatively short time in evolutionary terms, should be error-free or otherwise optimal (Marcus, 2008).

Simulation also seems in many cases to conflict with what we know about people's abilities to reason more generally. People have a limited working memory that can only keep track of a few items at a time; they are often insensitive to the addition or removal of objects from the world around them; they exhibit domain-general reasoning errors that conflict with the the laws of probability. All of these observations are in tension with the commitments of simulation, which assumes that people are tracking many objects at a time, in complicated ways, and without making basic logical errors. Our empirical investigations (Ludwin-Peery, Bramley, Davis, & Gureckis, 2020, 2021)

ELP: I cited the Psych Sci paper and the preprint version of the 3-experiment paper here (not the CogSci 2019 version), we could cite otherwise if someone wants

demonstrated that the behavioral results were more in line with what is known about cognitive science more generally, and inconsistent with simulation. Specifically:

1. Simulation requires that the cognitive process keep track of all the individual objects involved in the scenario. We found that, on the contrary, people find it hard to keep track of ten objects and indeed often fail to notice the addition or deletion of one or two objects.

2. Simulation requires time to advance in lock step across all parts of the scenario under consideration. We found that, in some cases, this kind of temporal coherence is not maintained.

3. The predictions of probabilistic simulation, however it is carried out, necessarily conform to the basic axioms of probability theory. We found that physical reasoning is subject to the same "conjunction fallacy" discovered by Tversky and Kahneman (1982, 1983). Moreover, these kinds of errors were made in a range of experimental settings: some of our tasks involved only one dynamic object, others involved many dynamic objects; some were presented as video

> clips, others presented as static images; some involved colliding balls, others involved towers of blocks. Further, this error persisted whether the question was asked in a straightforward, neutral, or even deliberately unusual manner.

Even sophisticated and experienced subjects, with time to think, can make elementary errors. Halloun and Hestenes (1985) found dismal level of performance in very simple problems involving force among college students, both on entrance to a freshman physics class, and after completing the course. Such errors on simple problems can be made, not only in a classroom setting with theoretical problems, but in a real-world setting, where the most serious possible consequences at stake, to a subject with both practical and theoretical knowledge. The psychologist Rebecca Lawson (2006, p. 1674) tells a remarkable personal anecdote:

> I regularly scuba dive, and on one weekend trip two friends and I decided to do a night dive. We needed our weight belts, which were on a boat in the harbor, so I offered to fetch them. I swam the few meters to the boat in my dry suit, clipped my own weight belt around my waist, held one weight belt in each hand, slipped overboard, and headed back to the ladder. To my surprise, I immediately sank to the bottom of the harbor. I flailed up and managed to gasp some air but could not move forward before I was dragged down again. Realizing that I was well on my way to drowning, I dumped the two loose weight belts and managed to reach the ladder. In retrospect, the mistake was shocking. I have spent years carrying lead weight belts around: Their heaviness is both perceptually highly salient and central to their function. I had tried to carry three people's weight belts, even though I knew that my own weight belt was adjusted to make me only just buoyant in a dry suit. Why, then, did I fail to anticipate what would happen when I jumped into the water carrying enough lead to sink me, no matter how hard I swam? My conceptual knowledge of weight belts let me understand what had happened after the event, but it was not until the physical constraints of the real world impinged that I accessed this information. Such post hoc explanations created after perceptual experiences allow us to learn from our mistakes, but otherwise may have little influence on our everyday actions.

A theory in which all physical reasoning was done by a veridical simulation that was accurate up to limits posed by perceptual noise and limits on working memory would offer little insight into why in many physical domains errors seem so pervasive.

4.2 Intermittent use of a correct physical theory

A challenge to a pure simulation account comes from the fact that people's accuracy can vary widely depending on the framing of a given problem, even if the underlying physics is essentially identical. Indeed, in some cases, two tasks that differ only in the desired form of the answer can yield entirely different results. For example, when subjects are asked to reason about the flight of the bob of a pendulum after it is cut, participants do well if they are asked to place a basket so as to catch the basket or to place a blade so the bob will reach a specified spot (K. A. Smith & Vul, 2013), but in the same study it was found that participants do poorly if they are asked to draw the trajectory of the bob.

To take another example, Battaglia et al. (2013) propose that reasoning about falling towers of blocks involves a full Newtonian theory of solid objects. However, in other settings governed by the same physical theory, people do very poorly. The principle underlying a balance beam, for instance, involves a rather simple subset of Newtonian physics; but it is well established that naïve subjects make systematic basic errors in reasoning about balance beams (Siegler, 1976) and we have shown (Marcus & Davis, 2013) that these errors cannot be accounted for using the model of imperfect perception in Battaglia et al. (2013). Note that a balance beam can be built as a kind of tower, and in fact will, with some probability, be built by the "random tower construction" algorithm that Battaglia et al. (2013) are using to generate test examples. Similarly, Newtonian theory of solid objects, and the specific physics engine used by Battaglia et al. (2013) also incorporates gyroscopic motion; but naïve subjects not cannot accurately predict the behavior of a gyroscope; indeed, they can hardly believe that a gyroscope functions as it does even when they directly experience it.

As a result, the predictive power of the "noisy Newton" theory is severely limited. It is not the case that people reliably use a noisy Newton method, or any kind of Newton-based method, for problems involving rigid solid objects. It is not even the case that they use reliably these methods for the specific problem of predicting the behavior of cutting pendulums. Until there is a characterization of which problems invoke “noisy Newton”, the theory does not make any general, testable predictions.

4.3 Implausibly rich tacit theories

As Sanborn et al. (2013) acknowledge, “Newtonian physics and intuitive physics might seem far apart”, given that the “The discovery of Newtonian physics was a major intellectual achievement and its principles remain difficult for people to learn explicitly.” Of course, we know that people can have tacit knowledge that can be very difficult to make explicit; linguists struggle to characterize language, while children manage to learn language readily without being able to articulate the underlying rules. Nonetheless, both the gap between people’s performances on physical reasoning tasks and Newtonian physics, and the inherent difficulty of characterizing the complexities of the real world in terms of scientific principles shed doubt on claims that human physical reasoning largely reflects correct scientific theories.

Consider, for example, the motions of bicycles; although they are familiar to most people, it turns out that the full theory of bicycle motion was not correctly understood until 2011 (Schwab & Meijaard, 2013).[2] A simple Newtonian physics engine will not deal properly with bicycles, because the feedback from the rider is critical. Certainly there were no bicycles during the principal period of human evolutionary adaptation; indeed there were neither wheels nor any controllable rolling objects of any kind. Perhaps human evolution found its way to an innate physics engine capable of detailed reasoning about systems in ways that only became useful in the late nineteenth century, but it seems unlikely. Instead, it seems more plausible to assume that our notion of bicycle mechanisms revolves around some sort of intuitive simplification that derives from experience.

To take another example, consider the process of cooking scrambled eggs. To what extent the physical chemistry that explains the change in physical characteristics and taste wrought by heat and stirring in turning a raw egg into a scrambled egg is currently understood we do not know, but it seems altogether implausible to suppose that naïve subjects have tacit knowledge of this physical chemistry. A subject who has never seen an egg being cooked would be very unlikely to be able to predict its behavior from first principles. Cooks seem to know “one-off” characteristics of eggs specifically, via experience, not an instance of any useful more general theory.

On the other hand, people are able to connect their very imperfect understanding of cooking eggs with their physical reasoning generally. They know, for example, that they can flick a piece into the air with a fork; and that if they turn the pan upside down, the eggs will fall out, even if they have never tried it. They can guess that if you mix blue ink into them they will probably turn blue; and that they can probably cook an ostrich egg the same way, but it will probably take longer. This ability to combine theories of very different levels of detail is very difficult to attain with physics engines, since a physics engine requires a precise theory; alternative reasoning techniques such as symbolic reasoning are much more flexible in that regard.

4.4 Simulation as the result, rather than the mechanism, of physical reasoning

In a number of cognitive studies of visualization or embodied simulation, it is clear on consideration that most of the physical reasoning involved must be carried out *before* the simulation can be constructed; and that therefore, the simulation cannot be the cognitive mechanism that supports the physical reasoning. For example, Zwaan and Taylor (2006) report an experiment in which subjects read one of the following two texts:

1) The carpenter turned the screw. The boards had been connected too tightly.
2) The carpenter turned the screw. The boards had been connected too loosely.

Subjects who read sentence (1) then find it easier to turn their hand in a counterclockwise direction, while subjects who read sentence (2) find it easier to turn their hand clockwise. But the connections between the texts and the directions of turning the screw themselves rest on a rather

[2] This despite the fact that a bicycle is a quite simple system; in terms of a stability analysis, a bicycle with two wheels on the ground consists of four rigid objects (the two wheels, the handlebar, and the frame) with six degrees of freedom, and numerous symmetries.

complex sequence of inference, combining reasoning about the physics with reasoning about the carpenter's purposes.

Similar considerations apply to non-physical visualizations as well. For instance, Moulton and Kosslyn (2009) discuss a hypothetical case of a groom who is considering telling a risqué wedding toast, visualizes the look of horror on the bride's face, and changes his mind. But how does the groom come to visualize his bride with a look of horror? The real work is beforehand; the cognitive process that generates the visualization must surely be drawing on a previous inference that the bride will be horrified based on knowledge of speech acts, ceremonial occasions, emotions, and so on. It is not plausible that there is a process that goes directly from the joke to the image of horror and then interprets the image to infer the emotion of the bride.

ELP: I like this section a lot but I worry some readers might not follow the argument. Could you expand it at all or go into more detail, do you think?

ED: I'm not seeing the problem, or how to make the point clearer. Gary, Todd, any thoughts?

4.5 Drawing the target around the arrow

One inadvertent consequence of the popularity of physics engine models in cognitive psychology is that, increasingly, the type of physical phenomena that are considered are those that work well with physics engines, rather than problems that are natural or ecologically valid.

For instance the papers (Carroll & Kemp, 2015), (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018), and (Ullman et al., 2018) all involve reasoning about disk-shaped objects that attract or repel each other at a distance and move on a surface with varying degrees of freedom. It is certainly possible to construct such objects by gluing fixed magnets around the edge of an air puck. Perhaps some high school lab has actually done that, though in practice this would not be a very good experimental design, as the motion would not be easy to measure and and the magnitude of the forces would probably not be easy to characterize. Certainly no one ever encounters such a thing in everyday life. Perceptible forces between separated mesoscopic objects of any kind are rare in quotidian environments; there are magnets and occasionally there are object with static electric charges. But this kind of physical set up, with its elegant, simple differential equation, is obviously catnip for a physics engine; so, to a cognitive psychologist who believes in mental physics engines, it becomes a natural domain of study.

Meanwhile, the kinds of situations that people actually encounter in everyday life, such as those quoted above from (Battaglia et al., 2013) — "a stack of dishes will topple,a branch will support a child's weight, a grocery bag is poorly packed and liable to tear or crush its contents, or a tool is firmly attached to a table or free to be lifted" — remain almost entirely unstudied. The stack of dishes has been replaced with the idealized and very different situation of a tower of rectangular blocks, aligned with each supported by exactly one other (figure 7). It is a safe bet that no physics engine has ever been built that can predict the breaking of a branch under a child's weight or the tearing of a grocery bag, and there is no evidence whatever that people would use a physics engine, or anything like one, in reasoning about these.

Likewise, as mentioned earlier, this line of research has focused exclusively on reasoning about how inanimate objects interact with one another. The questions of how inanimate objects behave in response to the actions of animate actors, which is surely much or most of the intuitive physical reasoning that people carry out, has been almost entirely ignored.

One could argue that, after all, people are giving answers to the problems with attractive and repulsive disks, so how they are finding those answers is a suitable matter for study. It may be interesting to study, but there is little reason to suppose that it shed any more light on how people do everyday physical reasoning, any more than studying the strategies that people use for memorizing random lists of words sheds light on how they understand language.

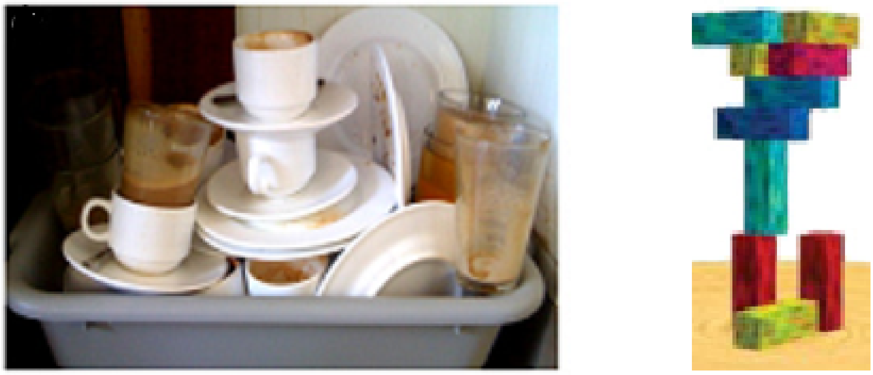


Figure 7: A stack of dishes vs. a tower of blocks. Both figures are from (Battaglia et al., 2013)

5 Non-veridical simulation

5.1 Partially specified models, supplemented with machinery borrowed from formal logic

A critic might respond to some of the points we have raised above by claiming that simulations do not have to be veridical; they can be cartoons or schemas that stand in for the world, rather than exact simulations of physical reality. SimCity doesn't need to have a one-to-one correspondence to real world financial institutions in order to be an orderly, internally coherent proxy that stands in for the real world and teaches children something about capitalism. Johnson-Laird (1983) even argues that mental models do not have to be fully specified; they can include unspecified elements. If our critique applied only to fully-specified veridical simulations, it would not prove very much.

Once one starts to introduce partially specified models, however, the distinction between models and alternatives such as inference using propositional logic starts to become hazy; the more flexible the kinds of partial specification allowed in the models, the hazier the distinction becomes. Johnson-Laird, for example, allows a symbol for "not". If one additionally allows symbols for "and" and "exists" and a scoping mechanism, the system becomes equivalent to first-order logic. At that point, it becomes unclear what distinguishes a "simulation"-based model from any other inference technique. The computational advantages of simulation and virtually all the empirical evidence in favor of simulation in cognition are grounded in an interpretation of simulation as an operation on fully specified descriptions of the world. If that is weakened, the theory becomes much more nebulous.

5.2 Non-realistic diagrams

People often find diagrams extremely useful in carrying out physical reasoning tasks. Physical situations that cannot be understood from any verbal description can be immediately apprehended given a picture, or, better yet, a video. The natural explanation of this is that useful geometric information that would be difficult to compute from a symbolic representation can be directly computed from a physical picture or from a picture-like internal representation. For example, (Chandrasekaran, Glasgow, & Narayanan, 1995, , p. xxii) write, "Diagrams preserve, or represent directly, locality information. A number of visual predicates are efficiently computed by the visual architecture from the above information, e.g. neighborhood relations, relative size, intersections, and so on. This ability makes certain types of inferences easy or relatively direct."

This is a plausible explanation of realistic drawings. An engineer or architect, for example, finds it useful to construct a scale drawing or model because all geometric relations can simply be measured, and any error that has been made in geometric calculation is immediately apparent. The problem with the explanation, however, is that people find non-veridical diagrams just as useful, in the same way, and in those cases the explanation is much more problematic. Consider, for example, Figure 8, from the Wikipedia article, "Redshift", which shows a star, a light beam, and an observer. This diagram is useful, but not veridical, given the extreme distortion of relative scale.

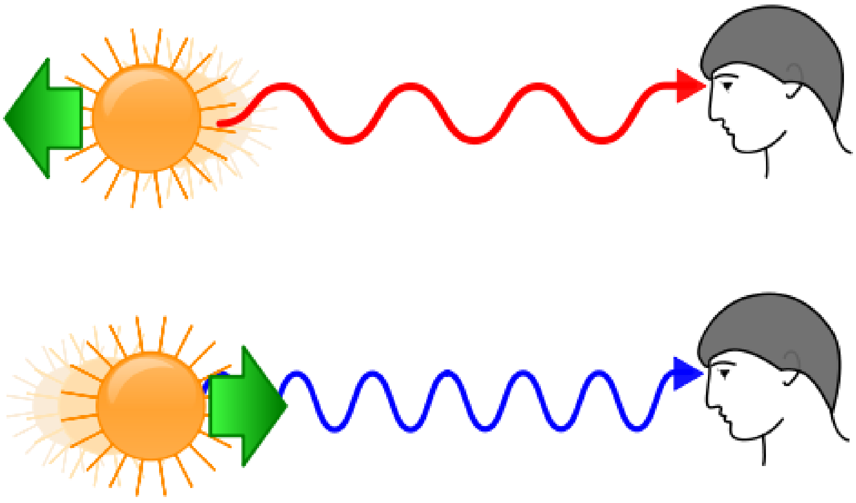


Figure 8: Star, Light, Observer. "Redshift and blueshift", by Aleš Tošovský from "Blueshift", Wikipedia. Image published under a Creative Commons 3.0 licence. https://commons.wikimedia.org/wiki/File:Redshift_blueshift.svg

The distance from the star to the observer is shown as equal to 4 wavelengths of red light; in reality, it is $4 \cdot 10^{22}$ wavelengths for the nearest star. Note that the wavelength is not purely an icon, like the big green arrows next to the star, or arbitrary, like the height of the wave in the transverse direction; the comparative wavelength of the blue and the red lights is meaningful and important to the point being illustrated. The scale of the star and the scale of the observer are likewise hugely distorted.

Similarly, Figure 9, from *The Feynman Lectures on Physics* (Feynman, Leighton, & Sands, 1964), illustrates molecules of gas in a piston.

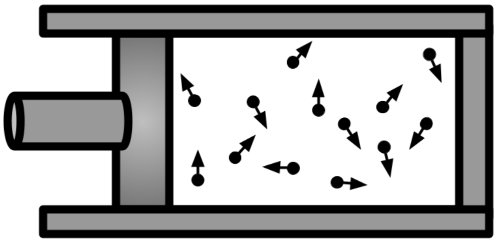


Figure 9: Gas molecules in a piston. Based on a figure from *The Feynman Lectures on Physics* (Feynman et al., 1964).

In the actual situation, there are of course about 10^{23} molecules of gas.

If diagrams are taken literally, serious problems start to arise in distinguishing conclusions that are true in the real situation from those that are merely artifacts of the diagram. For instance, if one took Figure 8 literally, one would conclude incorrectly that, since the light is shown as goes up and down on a path with a 4 inch amplitude, a 3x5 index card will sometimes fail to block one's view of an object directly ahead (even when the card is held directly in front of the eyes), and that the card would sometimes succeed in blocking it, even when the card is held 4 inches below the line of sight (Figure 5.2). There is nothing in Figure 8 to rule this out; it must be ruled out by reasoning that lies outside this picture. Similarly the student looking at Figure 9 is supposed to find it helpful in understanding that the motion of the molecules exert pressure on the inside of the piston, but they are not supposed to draw the invalid conclusion that the gas molecules collide much more often with the sides of the piston than with one another, which would be true in the situation depicted. Non-veridical simulations raise difficult problems of interpretation that can often only be solved using non-simulative reasoning. And if diagrams are not taken literally, the role of simulation in a larger theory once again becomes greatly diminished.

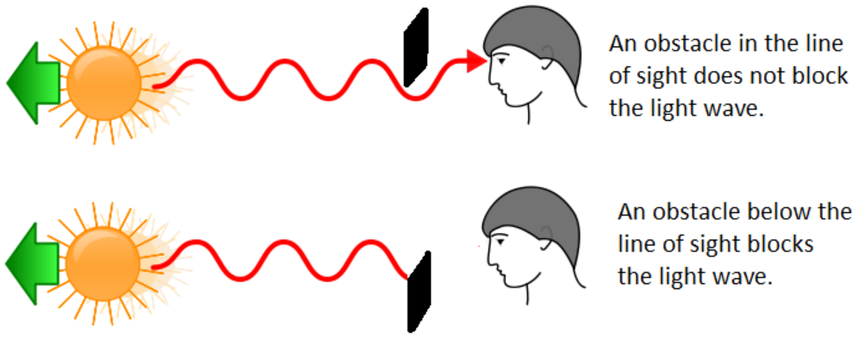

An obstacle in the line of sight does not block the light wave.
An obstacle below the line of sight blocks the light wave.

Figure 10: Invalid inferences from Figure 8. Image modified from "Redshift and blueshift", by Aleš Tošovský from "Blueshift", Wikipedia. Image published under a Creative Commons 3.0 licence. https://commons.wikimedia.org/wiki/File:Redshift_blueshift.svg

6 Perception, memory, and learning in physical reasoning

An alternative approach to modeling physical reasoning highlights the role of perception, memory, and learning. In this approach, people simply learn, from experience, to *see* whether a tower of blocks is stable or whether a cup of coffee will spill or where to place a bucket to catch a pendulum that has been cut, in the same way that they learn to see whether an image shows a cat or whether a video shows a ball going through a basketball net. For instance, Firestone and Scholl (2016) bring experimental evidence indicating that people learn to see whether a tower of blocks is stable as a direct perception; one learns to see whether a tower is stable as a visual feature in the same sense that one sees that a block is rectangular.

Such a model might simply output an answer — "the tower is unstable"; "place the bucket at $x = 2/3$"; that would not be a simulation-based system in any sense. Alternatively it might involve generating a full trajectory of a physical system; that could be viewed as a simulation-based system but not one based on a physics engine. For instance, Sanborn et al. (2013) propose that a "noisy Newton" model of collisions could be implemented using stored memories of observed collisions. When a new collision must be reasoned about, the reasoner would retrieve related stored memories and evaluate the new situation by a process of interpolation among the stored memories, taking into account uncertainty in the perception of the new situation. One could imagine that when an agent sees an open thermos of hot chicken soup knocked off a table edge, for example, they recall other cases where they have seen containers of liquid spill and predict that the results here will be similar.

With the explosive growth of machine learning technology, particularly deep learning and reinforcement learning, in the last decade and the concomitant interest in neuronal models of higher-level cognition, such approaches have become both more popular and more plausible. The development of actors, robotic or virtual, that use a combination of deep learning and reinforcement learning to learn to execute physical tasks is an active and growing area of research (Li, Leonardis, Bohg, & Fritz, 2019; Wu, Yan, Kurutach, Pinto, & Abbeel, 2019), though enormous challenges remain in applying in real-world robotics (Ibarz et al., 2021). Also related are the recent successful application in using deep learning to generate simulations for problems of scientific interest more quickly than conventional physics engines (Miyanawala & Jaiman, 2017; Raissi, Perdikaris, & Karniadakis, 2019).

In some respects, and for some kinds of problems, this approach to intuitive physical reasoning seems plausible and promising. One can well believe that this is an effective way to train a robot to carry objects from one place to another or to build a tower of blocks or to pour water from a pitcher into a glass. Certainly, the success of deep learning in physics applications such as fluid dynamics suggest that this kind of technology can at least sometimes deal with very complex physical systems. Since these methods do not rely on having detailed physical models, they might well be more robust than simulation-based methods on examples like the scrambled eggs (page 13) and the baseball pitcher (page 9) examples discussed above.

However, it is difficult to predict, or even to characterize, the capacities and limitations of this kind of learning technology trained on very large data sets in any new application. Notoriously, the output of the current generation of large language modeling is sometimes astonishing and sometimes risible. The key to using experience to solve new problems, in physical reasoning as elsewhere, is finding the proper generalizations. For example, to solve the gear problems of Schwartz and Black (1996) reliably, the system has to learn, implicitly, the rule that two interacting gears rotate in such a way that the contact point moves at equal speeds along the two boundaries or something similar, in addition to the more basic that gears rotate rigidly around a fixed axis. To solve the pulley problems of Hegarty (1992) the system must learn, implicitly, that the total length of the rope remains fixed; that the rope passes in a semi-circle around the pulleys and straight lines between pulleys; that each pulley rotate around an axis that can move vertically; and that a pulley rotates in the direction determined by the motion of the rope around it. (We are not, certainly, saying that the reasoner has to formulate these rules in any kind of symbolic or verbalizable form; just that correct predictions conform to these constraints.) Until a system is successfully deployed, there is simply no way to predict whether a cognitive process powered by deep reinforcement learning will arrive at the right generalizations in these and hundreds of other types of physical reasoning problems; what kind of data and how much data would be needed; and what kinds of anomalies, failures, and brittleness the trained system would exhibit.

Techniques of this kind are certainly much more sensitive to the particular form of the input than simulation methods. If the system has been trained on visual data, then it is not likely to be able to deal with information from other sources. Many of the significant material properties of the objects in a physical situation are not visible. It is straightforward to incorporate these in a simulation or in symbolic reasoning; incorporating them in a system trained by deep reinforcement learning is much more challenging.

Furthermore, some of the issues we have discussed as challenging for simulation methods are likely to be equally challenging for learning-based methods. Like simulation methods, it can be expected that the learning-based methods will do well dealing with closed systems over a short period of time. They are likely to have much more trouble dealing with radically incomplete information in an open world, such as our earlier example of a shirt inside a lost suitcase (page 10). (This is somewhat analogous to a system that can read an extended narrative and relate events at the end to events at the beginning. No existing AI reading system can do this.)

Another difficult problem is the need to merge multiple separate bodies of knowledge. Consider, for example, a subject who sees an open cup of coffee perched on top of an unstable tower of blocks. To predict what will happen, it is necessary to combine what has been learned of falling towers of blocks with what has been learned of spilling containers of liquids. In a knowledge-based system, such combinations are generally fairly straightforward, though certainly difficulties can arise; in a physics engine-based system, this can easy, if the engine already incorporates all interactions, or laborious, if it does not; but deep reinforcement learning does not, by any means, guarantee this kind of compositionality.

7 Conclusions

Simulation-based theories of physical reasoning are both too weak and too strong. They are too weak in that there are many forms of physical reasoning that people carry out where simulation is either extremely inefficient or entirely inapplicable. They are too strong, at least in the "noisy Newton" formulation, because in many tasks they predict much higher levels of performance than people exhibit; only in rare cases can the large, qualitative, fundamental, systematic errors that humans make be explained in terms of perceptual noise.

At this juncture, it is difficult or impossible to quantify what fraction of the physical reasoning in human cognition or in a general artificial intelligence is or could be carried out using simulation. We have argued, however, that the range of examples we have presented in this paper — many constructed as simple variants of problems considered in the pro-simulation literature — suggests that there are significant limits to the use of simulation. In particular, although we have suggested that simulation is effective for physical reasoning when the task is prediction, when complete information is available, when a reasonably high quality theory is available, and when the range of spatial or temporal scale involved is moderate, in many other cases simulation is to some extent problematic. In particular, physical reasoning often involves tasks other than prediction and information is often partial.

Moreover, even when optimal conditions hold, there are many cases in which it would appear that alternative non-simulative modes of reasoning are likely to be easier, faster, or more robust. Finally, setting up and interpreting a simulation requires modes of physical reasoning that are not themselves simulation. For all these reasons, we suggest that non-simulative forms of reasoning are not an optional extra in cognitive theories but are centrally important.

One reason that researchers have overstated the role of simulation in cognitive models is that in the majority of studies of physical reasoning in the psychological literature, subjects are asked to carry out task of prediction and in the overwhelming majority they are presented with complete specifications of the situation.[3] However, there is little reason to suppose that that at all reflects the frequency of these forms of reasoning in ecologically realistic settings; it may well result from artificial constraints that arise in setting up a controlled experiment.

We concur with Hegarty (2004) in seeing room for hybrid models that combine simulation with other techniques, such as knowledge-based inference. In our view, however, simulation may play a much less central role than Hegarty envisioned. In particular, in a general intelligence, human or artificial, practically any reasoning that involves simulation will probably involve some degree of non-simulative reasoning, to set up the simulation and to check that the answer is broadly reasonable. Outside of carefully constructed laboratory situations, and certain forms of specialized expert reasoning, we think it is relatively rare that simulation is used in isolation in human cognitive reasoning.

If there is indeed a physics engine in the head, it is likely to be only a small part of a larger system that encompasses a wide range of additional cognitive processes, such as learning, memory-based reasoning, causal reasoning, qualitative reasoning, rule-based heuristics, analogy, and abstraction.

Acknowledgments Thanks to Peter Battaglia, Benjamin Bergen, Neil Bramley, Thomas Bräunl, Alan Bundy, Chaz Firestone, Jonathan Goodman, Philip Johnson-Laird, Owen Lewis, Casey McGinley, Kevin Smith, Andrew Sundstrom, and Ed Vul for helpful discussions; and to Nils Thuerey for information about simulation for film CGI.

References

Abella, J., & Demircan, E. (2019). A multi-body simulation framework for live motion tracking and analysis within the unity environment. In *2019 16th international conference on ubiquitous robots (ur)* (pp. 654–659).

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and cognitive processes*, *18*(5-6), 513–562.

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.

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

Boeing, A., & Bräunl, T. (2007). Evaluation of real-time physics simulation systems. In *Proceedings of the 5th international conference on computer graphics and interactive techniques in australia and southeast asia* (pp. 281–288).

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

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

Carroll, C. D., & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition*, *139*, 130–153.

[3] There is also a tendency in the cognitive science literature to focus on idealized physical models that are characteristic of textbooks or engineered devices, such as balance beams, perfectly bouncing balls, and gears rather than those that occur in natural settings like camping trips or kitchens.

Chandrasekaran, B., Glasgow, J., & Narayanan, N. H. (1995). *Diagrammatic reasoning: Cognitive and computational perspectives.* Menlo Park, Calif.: AAAI Press.

Chung, S.-J., & Pollard, N. (2016). Predictable behavior during contact simulation: a comparison of selected physics engines. *Computer Animation and Virtual Worlds*, *27*(3-4), 262–270.

Collins, J., Howard, D., & Leitner, J. (2019). Quantifying the reality gap in robotic manipulation tasks. In *2019 international conference on robotics and automation (icra)* (pp. 6706–6712).

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062.

Davis, E. (2013). Qualitative spatial reasoning in interpreting text and narrative. *Spatial Cognition & Computation*, *13*(4), 264–294.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

DeRose, J., Wechsler, R. H., Tinker, J. L., Becker, M. R., Mao, Y.-Y., McClintock, T., … Zhai, Z. (2019). The aemulus project. i. numerical simulations for precision cosmology. *The Astrophysical Journal*, *875*(1), 69.

Farokhmanesh, M. (2021, march). Why game developers can't get a handle on doors. *The Verge*. Retrieved from https://www.theverge.com/platform/amp/22328169/game-development-doors-design-difficult

Feynman, R. P., Leighton, R. B., & Sands, M. (1964). *The feynman lectures on physics, vol. i: The new millennium edition: mainly mechanics, radiation, and heat* (Vol. 1). Addison-Wesley.

Fingas, J. (2018). Anyone can use nvidia's physics simulation engine. *Engadget*. Retrieved from https://www.engadget.com/2018/12/03/nvidia-physx-open-source/

Firestone, C., & Scholl, B. (2016). Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision*, *16*(12), 689–689.

Gerstenberg, T., & Goodman, N. (2012). Ping pong in church: Productive use of concepts in human probabilistic inference. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

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

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American journal of Physics*, *53*(11), 1043–1055.

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.

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 0278364920987859.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Johnston, B., & Williams, M. A. (2008). Comirit: Commonsense reasoning by integrating simulation and logic. *Frontiers in Artificial Intelligence and Applications*, *171*(1), 200–211. Retrieved from https://opus.lib.uts.edu.au/handle/10453/11426

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences*, *7*(8), 368–373.

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 *Cogsci.*

Kunze, L., & Beetz, M. (2017). Envisioning the qualitative effects of robot manipulation actions using simulation-based projections. *Artificial Intelligence*, *247*, 352–380.

Lambert, C. (2012). Twilight of the lecture. *Harvard magazine*, *114*(4), 23–27.

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & cognition*, *34*(8), 1667–1675.

Lawson, R. (2012). Mirrors, mirrors on the wall… the ubiquitous multiple reflection error. *Cognition*, *122*(1), 1–11.

Levillain, F., & Bonatti, L. L. (2011). A dissociation between judged causality and imagined locations in simple dynamic scenes. *Psychological science*, *22*(5), 674–681.

Li, W., Leonardis, A., Bohg, J., & Fritz, M. (2019). Learning manipulation under physics constraints with visual perception. *arXiv preprint arXiv:1904.09860*.

Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021). How do the object-file and physical-reasoning systems interact? evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, *125*, 101368.

Lu, L., Morse, M. J., Rahimian, A., Stadler, G., & Zorin, D. (2019). Scalable simulation of realistic volume fraction red blood cell flows through vascular networks. In *Proceedings of the international conference for high performance computing, networking, storage and analysis* (pp. 1–30).

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021, Jan). *Limits on simulation approaches in intuitive physics.* PsyArXiv. Retrieved from psyarxiv.com/xhzuc doi: 10.31234/osf.io/xhzuc

Marcus, G. (2008). *Kluge: The haphazard evolution of the human mind.* Houghton Mifflin Harcourt.

Marcus, G., & Davis, E. (2013). *Supplement to "how robust are probabilistic models of higher-level cognition.*

McCloskey, M. (1983). Naive theories of motion. *Mental models*, *14*(2), 299–324.

Miyanawala, T. P., & Jaiman, R. K. (2017). An efficient deep learning technique for the navier-stokes equations: Application to unsteady wake flow dynamics. *arXiv preprint arXiv:1710.09099*.

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

Mouret, J.-B., & Chatzilygeroudis, K. (2017). 20 years of reality gap: a few thoughts about simulators in evolutionary robotics. In *Proceedings of the genetic and evolutionary computation conference companion* (pp. 1121–1124).

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive psychology*, *22*(3), 342–373.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707.

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

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

Schwab, A., & Meijaard, J. (2013). A review on bicycle dynamics and rider control. *Vehicle System Dynamics*, *51*(7), 1059–1090.

Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. *Cognitive Psychology*, *38*(3), 433–464.

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

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

Sławiński, G., Niezgoda, T., Barnat, W., & Wojtkowski, M. (2013). Numerical analysis of the influence of blast wave on human body. *Journal of KONES*, *20*.

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

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

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., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In *Proceedings of the annual meeting of the cognitive science* *society* (Vol. 36).

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *science*, *332*(6033), 1054–1059.

Timperley, C. S., Afzal, A., Katz, D. S., Hernandez, J. M., & Le Goues, C. (2018). Crashing simulated planes is cheap: Can simulation detect robotics bugs early? In *2018 ieee 11th international conference on software testing, verification and validation (icst)* (pp. 331–342).

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

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.

Wang, T., Guo, Y., Shugrina, M., & Fidler, S. (2020). Unicon: Universal neural controller for physics-based character motion. *arXiv preprint arXiv:2011.15119*.

Watson, N. A., Kelly, M. F., Owen, I., & White, M. D. (2019). The aerodynamic effect of an oblique wind on helicopter recovery to the queen elizabeth class aircraft carrier. In *The vertical flight society-forum 75: The future of vertical flight-proceedings of the 75th annual forum and technology display.*

West, G., Ogden, M., Wallin, J., Sinkala, Z., & Smith, W. (2020). Optimizing numerical simulations of colliding galaxies. I. Fitness functions and optimization algorithms. *Research Notes of the AAS*, *4*(8), 136.

Won, I., Gross, S., & Firestone, C. (2021). Impossible somatisensation. Retrieved from https://psyarxiv.com/e5gy3/

Wu, Y., Yan, W., Kurutach, T., Pinto, L., & Abbeel, P. (2019). Learning to manipulate deformable objects without demonstrations. *arXiv preprint arXiv:1910.13439*.

Zickler, S., & Veloso, M. (2009). Tactics-based behavioural planning for goal-driven rigid body control. *Computer Graphics Forum*. Retrieved from http://www.cs.cmu.edu/~mmv/papers/09cgf-stefan.pdf

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*(1), 1.