Metareasoning and Mental Simulation

By
Jessica B. Hamrick

A Dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Psychology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Chair
Professor Tania Lombrozo
Professor Anca D. Dragan

Fall 2017

Abstract

Metareasoning and Mental Simulation

by

Jessica B. Hamrick

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Thomas L. Griffiths, Chair

At any given moment, the mind needs to decide *how* to think about *what*, and for *how long*. The mind's ability to manage itself is one of the hallmarks of human cognition, and these meta-level questions are crucially important to understanding how cognition is so fluid and flexible across so many situations. In this thesis, I investigate the problem of cognitive resource management by focusing in particular on the domain of *mental simulation*. Mental simulation is a phenomenon in which people can perceive and manipulate objects and scenes in their imagination in order to make decisions, predictions, and inferences about the world. Importantly, if thinking is computation, then mental simulation is one particular type of computation analogous to having a rich, forward model of the world.

Given access to such a model as rich and flexible as mental simulation, how should the mind use it? How does the mind infer anything from the outcomes of its simulations? How many resources should be allocated to running which simulations? When should such a rich forward model be used in the first place, in contrast to other types of computation such as heuristics or rules? Understanding the answers to these questions provides broad insight into people's meta-level reasoning because mental simulation is involved in almost every aspect of cognition, including perception, memory, planning, physical reasoning, language, social cognition, problem solving, scientific reasoning, and even creativity. Through a series of behavioral experiments combined with machine learning models, I show how people adaptively use their mental simulations to learn new things about the world; that they choose which simulations to run based on which they think will be more informative; and that they allocate their cognitive resources to spend less time on easy problems and more time on hard problems.

For Tobi, who is (and always will be) my favorite.

Contents

Listing of figures

Listing of tables

Acknowledgments

This thesis is a tapestry, whose threads were contributed by the many people in my life. Without the guidance of my mentors, the passion of my collaborators, the camaraderie of my peers, and the love of my family, I never would have accomplished all that I have.

I would like to begin by thanking my advisor, Tom Griffiths, for his neverending patience and encouragement. Tom believed in me and my research long before I did, and taught me to rephrase my questions in ways that were far deeper and interesting than I ever could have done on my own. I would also like to thank Josh Tenenbaum and Peter Battaglia, who have continued to be my mentors since I left MIT. My collaborations and conversations with Josh and Pete have profoundly shaped this thesis and my research, always for the better. I additionally owe a debt of gratitude to the members of my thesis committee, Tania Lombrozo and Anca Dragan. Tania helped me find a deep appreciation for the role of philosophy in cognitive science, and Anca reawakened my excitement and passion for artificial intelligence and robotics.

I have been incredibly fortunate to have so many friends and collaborators across so many areas of cognitive science, whose different perspectives and backgrounds always inspire me to think about problems in new ways. First and foremost I would like to thank M Pacer, who sees beauty in the details and who always asks the questions no one else thinks to ask.[1] M has been there for me since I arrived at Berkeley and has been a neverending source of support, inspiration, and true friendship. Both my research and my ability to think scientifically have been made better through our interactions. When I first started graduate school, I took so many assumptions for granted; M uncovered these assumptions, challenged them, and helped me rebuild my work to be stronger, more convincing, and more interesting. This thesis would be far weaker, and my life far less fulfilled, if not for M.

To David Bourgin, Thomas Langlois, Falk Lieder, and Stephan Meylan: I cannot believe we have almost come to the other side of our Ph.D.'s! I am so glad to have had you all with me on this journey. You all inspire me through the passion that you bring to your research and the way you think about problems; there are so many thoughts I would never have had were it not for our conversations and collaborations. I would also like to thank Joe Austerweil, Daphna Buchsbaum, Liz Bonawitz, Daniel Chada, Joseph Jay Williams, Luke Maurits, Anna Rafferty, Caren Walker, and Andrew Whalen, who made my first few years at Berkeley welcoming, exciting, and hopeful. The rest of the Computational Cognitive Science lab have continued to enrich my time at Berkeley: to Dawn Chen, Rachit Dubey, Nori Jacoby, Rachel Jansen, Alex Paxton, Josh Peterson, Avi Press, Aida Nematzadeh, and Daniel Reichman, I thank you. My gratitude also

[1] And who can predict my beer preferences better than anyone else, too.

goes to Jaime Fernández-Fisac and Chang Liu, with whom I have shared many late nights implementing awesome robot experiments and excitedly writing up the results before a deadline. To my other collaborators—Josh Abbott, Andy Ballard, Fred Callaway, Mike Frank, Monica Gates, Noah Goodman, Nicolas Heess, Dylan Hadfield-Menell, Razvan Pascanu, Dillon Plunkett, Tom Morgan, Kevin Smith, Jordan Suchow, Oriol Vinyals, and Ed Vul—it has been an honor to work with you and I could not have asked for kinder, more creative, or more intelligent colleagues. Thanks also goes to Todd Gureckis, Jay Martin, John McDonnell, Alex Rich, Doug Markant, and Anna Coenen: I have greatly enjoyed working with you on psiTurk and swapping drinks and ideas at CogSci. My other friends and colleagues outside of Berkeley have also had an enormous influence on me and this work; I would especially like to thank Kelsey Allen, Deborah Hanus, Andrea Tacchetti, and Alex Williams.

To Stephen Bailey, Will Jackson, and Elizabeth Seiver: I will sorely miss our weekly games of Exalted.

Like many Ph.D. students, I went through a period of doubting myself and my research, and was fortunate enough to have the support and encouragement of the IPython and Jupyter communities to raise me up during that time. I am especially indebted to Fernando Pérez, who made me feel welcome during my first IPython development meeting back in 2013, and who continues to make me feel appreciated every time we meet (even as my time spent contributing to the project has dwindled, particularly during the writing of this thesis). I am similarly grateful to Matthias Bussonnier and Thomas Kluyver, who initially helped me as mentors, and later hacked with me as good friends. I would also like to thank Brian Granger, who helped me start the nbgrader project; Min Ragan-Kelley and Kyle Kelley, who were foundational in helping me to get the "compmodels" JupyterHub deployment off the ground; and Carol Willing, who has been a role model and constant source of encouragement.

I truly owe everything to my family, who have never wavered in their love and support. My mother, Virginia Hamrick, has always been a pillar in my life and whose strength, grace, and kindness has simultaneously helped me through the hardest of times while inspiring me to always be a better and stronger person myself. My father, James Hamrick, never got to see me even start graduate school, but it is to him I owe my interest in AI and cognitive science. Were he still alive today, I know he would be as excited about my research as I am. Over the past five years, I have become even closer to my siblings, Ellie and David, and my sister-in-law, Rachel, even as we have lived on opposite sides of the country. They are fountains of endless good advice, and always leave me feeling happier and more certain of myself than before. My nephews, Charlie, Ben, and Matthew, are a source of light in my life whom I have watched grow over the past five years from little kids into curious, smart, and kind young men. The support of the rest of my family—and especially that of my grandparents, Peter and Sylvia Chandler—has given me the courage to pursue and achieve my dreams.

Last, but far from least, I would like to thank Tobias Pfaff, who loves me as I am—with all my quirks, ticks, and moments of indecision. Tobi has believed in me the whole way and has supported me even when I was on the verge of quitting. I truly believe that if it were not for his love and encouragement, I would have long ago abandoned this research and thesis; I am eternally grateful to him that I did not. Since we met, Tobi has brightened every day of my Ph.D. and has kindled in me a love and excitement for life and its possibilities that I did not know existed. I do not deserve his patience and kindness, and am thankful for it every day.

Funding Sources

None of the work in this thesis would have been possible without the generous financial support from a number of organizations. The Berkeley Fellowship I received from UC Berkeley, and the NSF Graduate Research Fellowship Program, gave me the means to live so that I could focus fully on my research. Grant number N00014-13-1-0341 from the Office of Naval Research allowed me to disseminate this work to the larger scientific community at a number of conferences, for which I am extremely grateful.

Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.

Albert Einstein

1 Introduction

In order to act efficiently and flexibly, people need to decide *what* to think about, *how* to think about it, and *how long* to think for. Such questions fall at the *meta-level* of cognition in that they are questions regarding "reasoning about reasoning", as opposed to questions about the reasoning process itself. To give a more concrete example, consider a game such as Angry Birds. In such games, people effortlessly decide the right way to represent the scene (i.e., should it be spatially represented in 2D or 3D, or propositionally represented as truth values?), the types of computations to perform to actually solve the task (i.e., should they rely on visual cues, engage in a process of logical deduction, or run a forward simulation of the scene's future?), and how much time to spend thinking (i.e., should they perform an uncertain computation multiple times, and at what point should they give up without having come to a solution?).

Questions about representation and processes have been fundamental to the field of cognitive psychology since its inception (Gardner, 1987). Indeed, as a testament to the importance of these notions, one may observe that some areas of cognitive psychology became embroiled in debates about the nature of representation for decades (e.g. Kosslyn, Thompson, & Ganis, 2006; Pylyshyn, 2002). In some ways, however, such debates have obfuscated the most important questions about cognition. Just knowing *what* representation people might be using does not tell us *why* they chose to use that representation, or *which* computations they choose to perform over that representation. If you were to learn that a software engineer used a computer program written in Python (the "representation") to assign category labels to pixel-based images, you would still not know very much about the actual solution to the problem.[1] It has been known for some time that understanding the process by which representations are manipulated is equally important as understanding the representation itself (Anderson, 1978), but even a focus on representation/process pairs misses out on the broader question of *why* that pair was chosen. If we truly want to understand how people decide what to think about, how to think about it, and how long to think for, a different approach is needed than a pure emphasis on representations themselves.

In this thesis, I assume a broad class of representations: forward, predictive models of the world which are broadly known as *mental simulation*. While mental simulation has been defined in many ways, and has been referred to in the context of many different empirical phenomena, its unifying theme is broadly the same: that a mental simulation is the repeated application of a model of what will happen next, either temporally or logically. Note that this definition of mental simulation does not make a strong commitment to the specific representation of the state space: it could be mental images (Kosslyn et al., 2006), logical propositions (Pylyshyn, 2002), or anywhere in between (e.g., mental models: Johnson-Laird and Yang (e.g., mental models: 2008); Khemlani, Mackiewicz, Bucciarelli, and Johnson-Laird (e.g., mental models: 2013)). For the purpose of this thesis, I typically assume the underlying representation is spatial (i.e., the positions, velocities, and geometries of objects) and that the simulation itself operates over a temporal scale. However, as I argue in Chapter 7, a general framework for understanding the meta-level decisions that people make need not be constrained by the assumption of spatial representation.

The aim of this thesis is twofold. First, I aim to demonstrate the power and flexibility of mental simulation, and to illustrate why it is a particularly interesting class of representations to focus on when thinking about the types of meta-level decisions that people make. Second, I examine several different meta-level questions that might be asked about mental simulation, and make the case for why the answers to these meta-level questions are important in understanding cognition more generally. In achieving both of these aims, I draw on the large existing literature on mental simulation and integrate it with sophisticated computational models based on advances in machine learning.

[1] Computer programs written in Python to assign category labels to pixel-based images are, today, quite common. For example, I could implement a deep neural network architecture such as that from Krizhevsky, Sutskever, and Hinton (2012) in a Python-based deep learning framework such as TensorFlow (Abadi et al., 2015). But I could also write a program to do image classification using Python that extracts SIFT features (Lowe, 1999) and then trains a Support Vector Machine (SVM) (Boser, Guyon, & Vapnik, 1992) to assign category labels. Simply knowing the program's inputs, outputs, and language tells us very little about what computations it performs.

1.1 Overview of the Thesis

In Chapter 2 ("Background"), I review what mental simulation is by giving an overview of the existing literature on mental simulation and forward models, with a particular emphasis on physical reasoning. In Chapter 3 ("Formalizing Mental Simulation"), I attempt to unify the phenomena described in the literature by giving a formal definition of mental simulation, and argue why it is useful to focus on simulation in particular, citing arguments from both the psychological and machine learning literatures. Finally, I explain the need for focusing on meta-level questions surrounding mental simulation.

In Chapter 4 ("Learning by Thinking"), I describe how mental simulation can be used not just for problem solving, planning, and prediction, but also for inferring unobservable properties of the world. I investigate a task involving inferring the mass ratio between different objects, and show that mental simulation combined with Bayesian inference can account for the inferences that people make. While these results do not explicitly concern the meta-level of reasoning, they do provide validity for the assumption that mental simulation can be thought of as a specific type of computational tool that can be reused across different tasks such as prediction and inference.

In Chapter 5 ("Selecting Computations"), I outline one approach to understanding how people decide *which* simulations to run. Specifically, I focus on the classic task of mental rotation (Shepard & Metzler, 1971). Even given the assumption that people are using mental simulation to imagine the shapes at different rotations, it remains unclear how people decide which rotations to imagine. I demonstrate that an active learning based approach to mental rotation is sufficient for producing behavior that is qualitatively similar to human behavior, while previously proposed models of mental rotation fail to capture these similarities.

In Chapter 6 ("Allocation of Cognitive Resources"), I switch from the question of which mental simulations people run to the question of *how many* mental simulations people run. In particular, if mental simulation is imperfect—as it often is, if only due to perceptual uncertainty—then running a single simulation does not give perfect information about the answer to a question. Running multiple simulations can provide a more precise estimate of this answer; yet at the same time, running too many simulations implies not acting and getting things done (Vul, Goodman, Griffiths, & Tenenbaum, 2014). I show that people adaptively change how many simulations they run depending on how difficult the task is, optimally solving the speed/accuracy trade-off imposed by trying to achieve a fixed level of accuracy across all potential decisions.

In Chapter 7 ("A Formal Framework for Modeling Mental Simulation"), I outline a theoretical framework for understanding meta-level decisions about mental simulation. This framework captures the results of Chapters 4-6 and provides a starting point for thinking about other classes of tasks as well, such as sequential decision making or inferring the preferences of other agents. Furthermore, this framework makes it clear which meta-level decisions we do not yet understand or have a good model for.

Finally, in Chapter 8 (“Conclusion”), I summarize the main contributions and results of the thesis and end with a discussion of the most exciting future directions.

Imagination will often carry us to worlds that never were. But without it, we go nowhere.

Carl Sagan

2 Background

Mental simulation has fascinated both psychologists and philosophers since at least as early as the ancient Greeks (Aristotle, 350 BCE).[1] This is, perhaps, partially due to the phenomenology of mental simulation (i.e., mental imagery) which is a distinctive feature of inner mental life for most people. It is perhaps also due to the fact that mental simulation is used so frequently and automatically across so many situations.[2] Yet, despite the long history of interest in mental simulation, its definition remains elusive.

Most researchers would probably agree that mental simulation is our ability to imagine interacting with objects, scenes, and other agents, even when the subject of those actions does not necessarily exist. So, I can mentally simulate picking up the cup in front of me, but I can also

[1] Aristotle discusses at length the nature of imagination, arguing that it is neither the same as sensation, nor the same as reasoning: "For imagination is different from either perceiving or discursive thinking, though it is not found without sensation, or judgement without it. That this activity is not the same kind of thinking as judgement is obvious. For imagining lies within our own power whenever we wish (e.g. we can call up a picture, as in the practice of mnemonics by the use of mental images), but in forming opinions we are not free: we cannot escape the alternative of falsehood or truth... that imagination is not sense is clear from the following considerations: Sense is either a faculty or an activity, e.g. sight or seeing: imagination takes place in the absence of both, as e.g. in dreams." (Book III, Part 3, Aristotle, 350 BCE)

[2] As a fun exercise, I encourage the reader to pay attention to how many times in a week people refer to the use of imagination, hypotheticals, or the "mind's eye". For example, you may notice your yoga teacher encouraging you to imagine that you are in a tranquil place during meditation at the beginning of class.

mentally simulate what it might be like to ride a unicorn. But there are other cognitive phenomena which do not fit neatly into this definition. I can imagine computing 3×5 by counting up by threes, five times—is that a mental simulation? When we move, our motor system uses forward predictions of the dynamics of the world in order to compute appropriate trajectories (Kawato, 1999)—are these forward simulations by the motor system also mental simulations? To nail down a more precise definition of mental simulation, I begin with a review of the literature on "simulation-like" phenomena in cognition. In Chapter 3 I attempt to unify these phenomena by a single definition.

2.1 Evidence for Mental Simulation

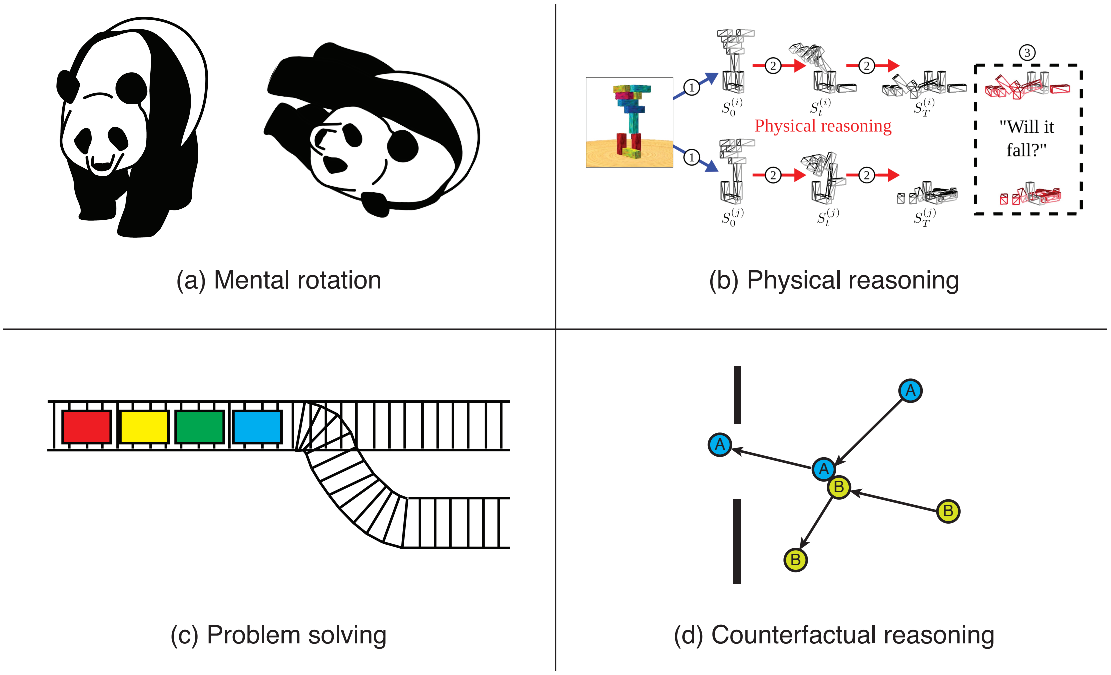
2.1.1 Mental imagery

The most well-known domain of mental simulation is likely that of *mental imagery*: the ability to picture something using one's "mind's eye" without actually perceiving it. Although mental imagery is commonly attributed to the visual domain, it can actually be found associated with every perceptual faculty. In this section, I briefly review the literatures surrounding the existence of mental imagery in each area of perception. For a much more detailed overview of the history and philosophical issues surrounding mental imagery, see Thomas (2016).

Here, I focus largely on the behavioral evidence for mental imagery, though I note there is a broad literature on the neuroscience of mental imagery as well. Briefly, evidence from the neuroscientific literature seem to converge on the fact that mental imagery—regardless of the domain—recruits areas of the brain that are associated with the corresponding perceptual faculty. For example, visual imagery recruits visual cortex (Kosslyn, 1988), and auditory imagery recruits auditory cortex (Zatorre, Halpern, & Ha, 2005). These results suggest that while mental imagery is not the *same* as perception, the two are deeply intertwined.

2.1.1.1 Visual imagery

The study of mental imagery as a phenomenon in its own right has been around since at least the beginning of the 20th century. For example, Perky (1910) ran an experiment in which participants were instructed to stare a blank screen an imagine an object such as a banana or leaf. Unbeknownst to the participants, the experimenters slowly projected an image of the object on the screen such that the projection was initially imperceptible, and only very slowly became perceptible. Remarkably, participants mistook the projection for their own mental imagery, suggesting



(a) Mental rotation

(b) Physical reasoning

(c) Problem solving

(d) Counterfactual reasoning

Figure 2.1: Examples of mental simulation tasks. (a) A variation on the classic mental rotation task by Shepard and Metzler (1971), in which the task is to determine whether the two objects are the same, or different. People solve this task by ``mentally rotating'' one object and comparing it to the other. (b) People reason about some physical phenomena, such as whether this tower of blocks will fall, by running mental simulations of the physical dynamics (Battaglia, Hamrick, & Tenenbaum, 2013). (c) An example of a problem solving task, in which people need to determine how to reverse the order of the cars (Khemlani et al., 2013). People solve this task by simulating the execution of an algorithm. (d) People make judgments of cause-and-effect based on counterfactual mental simulations (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014). For example, in this stimulus, people might judge that B caused A to go in the hole, because it would not have gone in had they not collided.

that there are deep ties between visual perception and mental imagery. The effect demonstrated in the experiment, now referred to as the *Perky effect*, is often used to demonstrate engagement of the perceptual system by showing degraded performance on some task that is hypothesized to use mental imagery (e.g. Bergen, Lindsay, Matlock, & Narayanan, 2007).

With the rise of behaviorism, the study of mental imagery declined through the middle part of the 20th century, and was only revived in the 1970s with the seminal work of Shepard and Metzler (1971). As illustrated in Figure 2.1a, Shepard and Metzler (1971) showed participants two images of similar objects and asked them to determine whether the objects were the same or different.

The striking finding was that, as the angle of rotation between the two objects increased, people's response times *also* increased. The authors concluded "that to make the required comparison [the participants] first had to imagine one object as rotated into the same orientation as the other" (Shepard & Metzler, 1971, p. 701).

The result from Shepard and Metzler (1971) spurred a huge body of research on mental imagery. Researchers subsequently investigated the specific processes underlying mental rotation (e.g. Just & Carpenter, 1976), mental rotations of other types of objects (e.g. Cooper, 1975), and other spatial tasks such as visually scanning between points in an image (Kosslyn, Ball, & Reiser, 1978). There are a number of phenomenon that are now considered to be classic, canonical examples of mental imagery (Kosslyn et al., 2006): generating novel and/or spontaneous images, scanning between two points in an image (Kosslyn et al., 1978), changing the size of an image (Bundesen & Larsen, 1975; Sekuler & Nash, 1972), and of course, mental rotation (Shepard & Cooper, 1982; Shepard & Metzler, 1971). Researchers have also shown compelling demonstrations of retinotopic activation in the visual areas of the brain when participants are asked to use visual imagery (e.g. Slotnick, Thompson, & Kosslyn, 2005). For more in-depth reviews of this literature, see Kosslyn, Pinker, Smith, and Shwartz (1979); Kosslyn et al. (2006). The research on mental imagery also ignited one of the most famous debates in cognitive psychology about the nature of mental representation and whether mental imagery is reflective of depictive or propositional representations. I discuss this debate further in Section 2.1.1.5.

Given the range of different types of imagery (discussed further in the next sections), it should not be surprising that those who have lost the capacity for vision still have imagery, even if it might not be explicitly visual in nature (Strauss Marmor & Zaback, 1976). People who have become blind later in life still possess visual imagery, though the nature of this imagery appears to change the longer an individual has been blind (Hollins, 1985). Congenitally blind individuals also have imagery, though it is perhaps more appropriate to call it *spatial* imagery rather than visual imagery (Arditi, Holtzman, & Kosslyn, 1988; Farah, Hammond, Levine, & Calvanio, 1988).

2.1.1.2 Auditory imagery

After visual imagery, auditory imagery is perhaps the second-most well-known form of mental imagery. Indeed, most people have had the experience of having an "earworm" (otherwise known as having a song stuck in one's head). Some evidence suggests that auditory imagery is separate from visual imagery, in that people with and without musical training exhibit different strengths of auditory imagery but do not differ in terms of their visual imagery (Aleman, Nieuwenstein, Böcker, & de Haan, 2000). But, there are also clear indications that auditory imagery interacts both with visual and motor imagery (see Zatorre et al., 2005, for a discussion). This suggests, perhaps, that we learn separate models for different perceptual domains, but that those models are linked particularly in cases where cross-modal perception already occurs. For example, there is evidence for activity in the motor areas of the brain when trained piano players hear a piece of music that they know how to play (Haueisen & Knösche, 2001).

2.1.1.3 Motor imagery

If there is one modality where we would *a priori* expect to find predictive models of the world, it is the motor system. For example, in order to catch a falling object, our motor system needs to be able to predict the future location of that object and plan an appropriate motion trajectory to that location (Wolpert & Flanagan, 2001). As would be expected, there is indeed strong evidence for forward predictive models of both dynamics and kinematics in the motor system (Kawato, 1999; Tong, Wolpert, & Flanagan, 2002). There is additionally evidence for inverse models (which map from initial states and desired future states to actions) and that these inverse models work in tandem with the forward models (Wolpert & Kawato, 1998).

Some researchers have argued that motor control does not require forward models (Gigerenzer & Brighton, 2009). For example, in the case of the "gaze heuristic", a ball may be caught by running towards it and keeping the angle of one's gaze toward the ball constant (McLeod, Reed, & Dienes, 2006). However, it is unclear how motor behavior would be produced in novel environments with novel objects—when heuristics do not hold—without some predictive mechanism. It is not inconceivable that the mind relies both on predictive models of control, as well as learned policies (of which the gaze heuristic is an example, see Belousov, Neumann, Rothkopf, & Peters, 2016).

The phenomenon of *motor imagery* is related to forward and inverse models in the motor system, in that those models are also engaged during motor imagery. According to Jeannerod (1995), "a motor image is a conscious motor representation" (Jeannerod, 1995, p. 1419), and indeed, there are many parallels between simulated motor action and real action (e.g. Gardony, Taylor, & Brunye, 2014; Parsons, 1994). Most striking is the phenomenon of "mental practice" (Feltz & Landers, 1983), in which mental simulation of movements can improve actual movement.

2.1.1.4 *Gustatory and olfactory imagery*

Although gustatory and olfactory imagery are arguably weaker forms of imagery than visual, auditory, and motor imagery, there is nonetheless clear evidence that they exist and that they have both behavioral and physiological signatures. For example, researchers have demonstrated that gustatory imagery increases salivation over non-gustatory imagery (K. D. White, 1978) and that the strength of food cravings is correlated with vividness of gustatory imagery (Tiggemann & Kemps, 2005). Olfactory imagery increases the rate at which participants sniff (over that measured during visual imagery), and people are likely to sniff more when imagining a pleasant odor rather than an unpleasant odor (Bensafi et al., 2003). For a review of the literature on olfactory imagery, see Stevenson and Case (2005).

2.1.1.5 *Challenges to imagery*

One of the most well-known debates in the psychology literature is the mental imagery debate throughout the 1970s, 1980s, and 1990s regarding the format of the representations underlying mental imagery. There were two main sides to this debate, which I summarize briefly below. I do not go further into the details but point interested readers to Pylyshyn (2002) and Kosslyn et al. (2006) for more recent reviews of the relevant literature.

On one side, we have the "depictive" argument (advanced largely by Steven Kosslyn), which posited that the format of mental images is *depictive*—i.e., an image. A depictive format might be a literal image (i.e., a drawing) or it might be a functional image (i.e., an array of pixels in a computer); in either case, the representation is tied to a particular modality (such as vision) and tends to mimic the properties of the relevant perceptual modality. Specifically, "depictive representations make explicit and accessible all aspects of shape and the relations between shape and other perceptual qualities (such as color and texture), as well as the spatial relations among each point" (Kosslyn et al., 2006, p. 14).

On the other side, we have the "propositional" argument (advanced largely by Zenon Pylyshyn), which posited that the format of mental images is *propositional* and symbolic, akin to to a LISP computer program. Importantly, Pylyshyn argues that there is nothing about the depictive representation itself that suggests any of the classic mental imagery results. He argues that these results only occur via the manipulation of that representation, which can just as easily be implemented symbolically as it can any other way. As he states, "what makes cells in a matrix appear to be locations with properties such as adjacency, between-ness, alignment, distance, and so on, is not any property of the matrix, nor even of the way that this data structure must be used.

There is no sense in which any pairs of cells is special and so there is no natural sense in which some pairs of cells are 'adjacent,' including a sense that derives from how they must be accessed" (Pylyshyn, 2002, p. 167).

As it stands, the imagery debate has never been officially "resolved", though most researchers seem to accept the depictive side of the argument. It is my opinion that, whether or not the underlying representations of mental imagery are depictive or propositional (or something else entirely), they are *spatial* and sensitive to the *causal structure* of the world—and it is these properties that are most important. The precise representation itself is somewhat secondary; in fact, as argued by Anderson (1978), it is not actually possible to fully disentangle these representations from behavioral data alone. The more relevant question, in my opinion, is to what extent they capture the underlying causal processes of the world (such as that an object's trajectory through space in time is piecewise smooth). The representation that best captures the causal structure of the perceptual world is spatial (meaning that it easily affords the computation of things such as distances between objects, their size, texture, color, and so on). Whether a spatial representation need be depictive (like a matrix of pixel values) or propositional is less relevant, because both can represent spatial information equally well.[3] What matters is not the format of the representation itself but the functional and computational properties of the representation.

A perhaps more serious challenge to the theory of mental imagery is evidence that some people do not have imagery at all (Galton, 1880). Imagery abilities seemingly lie on a continuum, with roughly 30% of individuals possessing exceptionally vivid and lifelike mental images, and with 2-5% reporting dim or even nonexistant images (Faw, 2009). This condition, which has only recently been named in the literature as *aphantasia*[4] (Zeman, Dewar, & Della Sala, 2015) seems to exist in both congenital and clinical forms, with varying levels of impact on peoples' lives. In some cases, aphantasia can have a noticeable and substantial effect, as in this anecdote from Faw (2009): "another psychologist…told me about how she had lost visual imaging ability from an auto accident in the past year. For months she found it hard to understand some of what she heard because she could not convert the words into pictures. Over a 6 months time she learned to encode all she heard through auditory imagery and regained comprehension without visual imagery" (Faw, 2009, p. 18). In other cases, people with aphantasia can go decades without realizing it (B. Ross, 2016) and they may not have any noticeable impairments (e.g. Zeman et al., 2010). It may be the case that people with aphantasia use alternate spatial reasoning strategies that

[3] In fact, a matrix of pixel values implemented in any modern digital computer *is* implemented symbolically under the hood.

[4] For a particularly revealing and personal account of aphantasia, I recommend B. Ross (2016).

do not require an explicit imagery or simulation component, and that in particular, congenital aphantasia is not noticed because those people have always used alternate strategies. Alternately, it is possible that people with aphantasia are perfectly adept at running simulations, but that the *phenomenology* of those simulations is absent and that they must retrieve the results of their simulations through some alternate means beyond imagery.

2.1.2 Memory and imagination

Evidence from the neuroscience literature suggests deep ties between episodic memory and episodic simulation. It has long been known that our memories do not encode every detail of an event, but rather that the brain "fills in" relevant missing details (Bartlett, 1932) in a reconstructive process of recall. This reconstruction can be thought of as a type of mental simulation (Kent & Lamberts, 2008) and suggests that the process of constructing a novel item—such as imagining a future event—might be related. Indeed, neuroscientific work in this area has established a strong relationship between episodic memory and future imagination (Hassabis & Maguire, 2009; Schacter et al., 2012). This suggests that, far from being solely a high-level cognitive phenomena, mental simulation is deeply tied to some of the mind's most basic processes, such as memory.

2.1.3 Language

While we frequently engage in mental simulation to imagine future or alternate possibilities, we also communicate our imagined scenarios to others in the form of poetry, prose, oral storytelling, and even in everyday conversations. It does not seem a stretch to claim, then, that the concepts engaged by language form a sort of mental simulation that allows us to envision these possibilities. This position has been argued most famously by Barsalou (1999), who suggested that the conceptual knowledge needed to comprehend language is composed of "perceptual symbols", which correspond to simulations of sensory and perceptual experience. In a related account, Zwaan and Radvansky (1998) and Zwaan (1999) proposed the notion of "situation models" in language comprehension, in which people construct a mental simulation of the scene being described. Further work has found evidence tying together situation models and perceptual simulations, such as faster response times for recognizing objects in a configuration consistent with the described scene (Zwaan, Stanfield, & Yaxley, 2002) or for determining the consistency of a sentence describing fast motion as opposed to slow motion (Matlock, 2004). Similarly, Bergen et al. (2007) found that spatial language can produce visual interference effects while performing a visual categorization task, and Dils and Boroditsky (2010) reported a visual motion aftereffect from spatial language comprehension. Thus, there seems to be a clear link between language comprehension and some form of mental simulation, though Willems, Toni, Hagoort, and Casasanto (2010) caution that implicit mental simulation of the form engaged in language comprehension is different from explicit mental imagery.

2.1.4 Intuitive physics and intuitive psychology

Psychologists have long studied our "intuitive theories" about both the physical world and the social world, and have appealed to the notion of simulation to explain how these intuitive theories work. For example, Battaglia et al. (2013) argued that people run simulations from an "intuitive physics engine" to make predictions about physical scenes such as towers of building blocks (Figure 2.1b); this area is discussed in further detail in Section 2.2. Similarly, others have argued for a simulation theory of mind-reading (Gallese & Goldman, 1998; Goldman, 1992; Gordon, 1992), though this theory has been hotly contested by advocates of the "theory theory" (Gopnik & Wellman, 1992; Saxe, 2005). As will become more obvious in Chapter 3, both the simulation theory and the theory theory can both be viewed as forms of mental simulation. In this respect, the term "simulation theory" is misleading because *both* the simulation theory and the theory theory imply running mental simulations: the distinction is in the form of those simulations. While the simulation theory holds that we use our own decision making neural circuitry to simulate what another agent would do (i.e., mirror neurons, see Gallese & Goldman, 1998), the theory theory argues that we construct a mental model of other agents' behavior. In both cases, we have access to some "theory of mind" that allows us to run simulations of what other agents may do, want, or believe.

2.1.5 Problem solving

Mental simulation is an important component of our ability to actually solve problems and plan. Problem solving itself has long been thought of as a kind of directed search through a difficult problem space (Anderson, 2015; Newell, Shaw, & Simon, 1958), which can be thought of as a type of mental simulation. Much early work explicitly focusing on mental simulation was devoted to how people construct "mental models" about complex systems (such as a calculator) and form plans for interacting with those systems (Gentner & Stevens, 1983). More recently, Khemlani et al. (2013) showed that people use mental simulation to reason through puzzles that have very algorithmic-like solutions, such as determining how to reverse the order of a set of objects (Figure 2.1c). Other work has shown that mental simulation underlies people's ability to perform arithmetic calculations with a "mental abacus" (Frank & Barner, 2011; Stigler, 1984), and to reason logically and perform deductive inferences (Johnson-Laird, 2012).

2.1.6 Thought experiments and counterfactual reasoning

It has also been argued that mental simulation is a key component of thought experiments and scientific reasoning. Gendler (1998) argues that thought experiments are a legitimate form of argument precisely *because* they provide access to a form of knowledge—through mental simulation—that would not otherwise be available through deductive reasoning alone. For example, Gendler (1998) recounts Galileo's famous thought experiment refuting the idea that objects of different mass fall at different speeds and reconstructs it as a deductive argument. She shows that the deductive argument draws its power not from the deduction, but from the believability of the premises; this believability comes from tacit knowledge that it is engaged by the thought experiment. Similarly, Clement (2009) and Trickett and Trafton (2007) show empirically that people spontaneously engage such knowledge by running new thought experiments and mental simulations when trying to understand a mechanical or scientific system.

A domain closely related to thought experiments is that of causal and counterfactual thinking, which also seems to engage mental simulation (Gerstenberg et al., 2014; Gerstenberg, Peterson, Goodman, & Tenenbaum, 2017; Walker & Gopnik, 2013). Counterfactual reasoning typically involves reasoning about different possible futures and what might cause those different futures to occur. For example, Gerstenberg et al. (2014) has shown that people run counterfactual simulations of colliding balls to determine whether one ball caused another to go in a goal (Figure 2.1d).

2.2 Mental Simulation in Physical Reasoning

This thesis focuses in particular on mental simulation in the domain of physical reasoning. The justification for this is that, unlike many other domains of simulation (language, theory of mind, etc.) we have a good understanding of the underlying physical principles of the world. Thus, using these principles we can construct ground-truth baselines to compare against, whereas it is more difficult to do so in other domains. This does not imply a commitment to the notion that people have a perfectly accurate intuitive physics (though as discussed below, it often comes close) but rather makes modeling mental simulation more tractable.

In this section, I describe the main approaches to understanding intuitive physical reasoning. Historically, there have been two parallel lines of research: one focusing on the limits of people's physical inferences, and one exploring the relationship between physical knowledge and mental simulation. More recent work has explicitly tried to unify these two lines of research and has brought a variety of new tasks and models to bear on the question of how people reason about the physical world.

2.2.1 Errors in physical reasoning

The terms "intuitive physics" and "naïve physics" were used originally by McCloskey and colleagues to describe errors that a number of people seemed to make when reasoning about they physical world (McCloskey, 1983). For example, Caramazza, McCloskey, and Green (1981) and McCloskey and Kohl (1983) showed that many people made a variety of errors when reasoning about the motion of several pendulum-like systems and also in the motion of balls in curved tubes. McCloskey, Washburn, and Felch (1983) similarly showed errors in reasoning about the path of a ball that rolls off a cliff or that is dropped while walking. The conclusions from this line of research were that people rely on an erroneous form of physical knowledge termed *impetus theory* after the Aristotelean notion that objects are imparted "impetus" when thrown, which propels them forward until the impetus dissipates.

Other research on physical reasoning, particularly in the domain of inferring mass from collision events, argued either that people rely on biased heuristics (Gilden & Proffitt, 1989a, 1989b, 1994) or that they "directly perceive" properties of the physical world (Runeson, 1977; Runeson, Juslin, & Olsson, 2000). In either case, the argument was that peoples' ability to make inferences about properties in the physical world from visual information is at best limited to a single dimension, if not also strongly biased. Although the heuristics-versus-direct-perception debate did not connect directly to McCloskey's work on intuitive physics, the conclusion was largely the same: people perform poorly when asked to reason about the physical world.[5]

2.2.2 Mental models

In parallel to the work focusing on errors in people's physical reasoning abilities, another line of work investigated the extent to which people could use mental simulation to reason about the physical world. In particular, this research focused on the idea that people construct "mental models" of physical objects and then simulate from those models (Gentner & Stevens, 1983).

[5] McCloskey and colleagues did admit the possibility of multiple sources of physical knowledge, however: "it does seem wise to recognize that a person may possess a perceptual appreciation of the natural dynamics of physical events, yet be unable to draw upon this knowledge when asked to conceptualize an event's outcome in a representational context" (Kaiser, Proffitt, & McCloskey, 1985, p. 539)

Research in this domain often worked under the assumption that mental models are not perfect models of the world, but that they are still rich enough to draw interesting predictions and inferences (Hegarty, 2004). For example, Hegarty (1992) showed how people can reason about mechanical systems such as a series of pulleys. Other work has shown how mental simulation can indeed lead to accurate inferences about physical scenes, but that people do not necessarily rely on mental simulation by default (Schwartz, 1999; Schwartz & Black, 1999).

2.2.3 Qualitative physics

A closely related area of research to the mental models literature is the qualitative and naïve physics literature from artificial intelligence (de Kleer & Brown, 1984; Forbus, 2011; Hayes, 1979, 1985; Kuipers, 1986). Work in qualitative physics has attempted to show how inferences can be drawn only from qualitative assertions—for example, determining whether a ball on the left side of the screen could collide with another ball on the right side of the screen (Forbus, 1983). Other work in qualitative physics has focused on commonsense reasoning of liquids (Davis, 2008), boxes (Davis, 2010), and containers more generally (Davis, Marcus, & Chen, 2013). While much of the work in qualitative physics has focused on logic-based reasoning, some has explicitly incorporated notions of simulation (Forbus, 1983; Gardin & Meltzer, 1989). However, while qualitative physics research has often been inspired by the notion of cognitive plausibility, it is usually not explicitly compared to human behavior.

2.2.4 Representational momentum

Another line of research in cognitive science which focused on peoples' physical knowledge focused on the phenomenon of *representational momentum* or *displacement*, in which people's memory for the location of an object is distorted in the direction of anticipated motion (Freyd & Finke, 1984). For example, people will misremember an object floating in the air as being lower than it actually is, perhaps due to expectations about gravity (Freyd, Pantzer, & Cheng, 1988). Additional research has revealed displacement effects for a wide range of physical scenarios, including circular motion, friction, and bouncing (see Hubbard, 2005, for a review). Related research has further found evidence for the automatic computation of implied motion when looking at still photographs (Winawer, Huk, & Boroditsky, 2008).

2.2.5 The “noisy Newton” hypothesis

There has been a recent resurgence of interest in cognitive science in understanding human physical reasoning, which both attempts to address classical physical reasoning errors and which formalizes the notion of simulating from a mental model. Termed the “noisy Newton” hypothesis, this body of research hypothesizes that people have an approximate knowledge of Newtonian physics which is instantiated through noisy physical simulation (Battaglia et al., 2013; Sanborn, Mansinghka, & Griffiths, 2013; Teglas et al., 2011). Thus, this hypothesis departs from the previous line of research which assumed that people have limited and poor knowledge of physical dynamics: instead, it shows how these errors might occur as a result of having rich but imperfect knowledge of Newtonian mechanics (e.g. Sanborn et al., 2013) or even that such errors do not occur when people are placed in a more naturalistic task (e.g. Smith, Battaglia, & Vul, 2013). The “noisy Newton” hypothesis also addresses what it means to “simulate” from a mental model by assuming that people represent the three-dimensional structure of the physical world and can approximately simulate the physical dynamics of that structure (Battaglia et al., 2013).[6]

“Noisy Newton” models have been used to explain human reasoning in an ever-growing number of tasks. For example, such models have been used to explain how people make predictions in complex physical scenes such as towers of blocks (Battaglia et al., 2013; Gerstenberg, Zhou, Smith, & Tenenbaum, 2017); how people reason about the behavior of fluids (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Kubricht et al., 2016) and other substances (Kubricht et al., 2017); how people reason counterfactually (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Gerstenberg et al., 2014); and how people engage in physical problem solving (Yildirim, Gerstenberg, Saeed, Toussaint, & Tenenbaum, 2017). Researchers have investigated in detail where the noise in peoples’ physical reasoning comes from (Smith & Vul, 2013); how people make physical predictions over time (Smith, Dechter, Tenenbaum, & Vul, 2013; Smith, Peres, Vul, & Tenenbaum, 2017); what neural systems might support such physical reasoning (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016), and the development of intuitive physics in children (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). For a recent review of this literature, see Kubricht et al. (2017).

[6] I note, however, that the models that have been researched so far do not span the whole space of possible simulation-based models. The existing simulation models typically assume a reasonably detailed representation of the scene, and that people simulate forward using a uniform time resolution. Both of these assumptions may be challenged, and there is evidence that they likely do not hold in all situations. For example, Hegarty (1992) showed that people can break down a physical system into parts and simulate the parts separately; Levillain and Bonatti (2011) showed that the error in people’s mental simulations does not always scale perfectly with time.

2.3 Summary

In this chapter, I have discussed a wide range of literature on mental simulation, ranging from mental imagery to thought experiments. The scope of mental simulation is truly vast, with phenomena that seem “simulation-like” appearing in nearly every aspect of cognition. These different forms of mental simulation are surely not the same. However, as I will argue in Chapter 3, they may be viewed through the same definitional framework. While my definition of mental simulation is broad, it allows us to see more clearly the similarities and differences between various forms of mental simulation, and makes it clear that the most interesting questions about mental simulation are those regarding *how* it is used.

If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.

Craik (1943, p. 61)

3 Formalizing Mental Simulation

Given the diverse array of phenomena associated with mental simulation, one might feel at a loss for defining exactly what mental simulation is. What definition could span the full range from mental imagery of all modalities, to memory and imagination, to language comprehension, to intuitive physics, to theory of mind, and to thought experiments? I would like to make the claim that we *can* find a definition that fits all of empirical results described previously, though that definition is necessarily broad. Despite this, I believe it is more useful to have a broad yet formal definition rather than no definition at all.

3.1 A Definition for Mental Simulation

The vast majority of research on mental simulation to date has focused on establishing the existence of mental simulation, describing what it can be used for (e.g., that it can be used to reason about a physical system), or characterizing its neural signatures. But, *why* do we even have mental imagery in the first place? To answer this question, it is helpful to look at mental simulation through the lens of Marr's levels of analysis (Marr, 1982) and Anderson's rational analysis (Anderson, 1990). I argue that mental simulation itself is a algorithmic-level phenomenon, but there are a broad class of computational-level problems that are best solved by mental simulation.

At the computational-level, there are a class of problems that involve being able to form predictions about future or hypothetical unobserved events. For example, we might want to predict what someone will do next; what will happen if I throw an object; or what the world would look like if I had done something differently. Such problems are best solved by having access to a rich, generative model of the world because such models are flexible, generalizable, and allow us to make strong inferences from small amounts of data (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Importantly, mental simulation also allows us to reason about phenomena in the world *even if those things have not actually occurred*, which is what makes mental simulation so important for counterfactual reasoning, thought experiments, and creative thinking. To put this discussion more succinctly: the best way to reason about possible futures of the world that we have not observed is to build a *model* of the world and simulate from it.

Using terminology from the machine learning and reinforcement learning literatures in computer science, I propose the following definition of mental simulation that solves the computational-level problem of forming predictions about unobserved events (Marr, 1982):[1]

> **All processes of mental simulation can be formally described either as the use of a transition function or an observation function.**

A *transition function* is a process that transforms one state of the world to another state of the world. For example, consider a classic arcade game such as Pong, in which you may move a paddle up and down in an attempt to bounce a ball off of it. The transition function for this game governs how the paddle moves as a result of the action taken (e.g., if the "up" action is taken, the paddle moves upwards by some amount) and also governs how the ball behaves (e.g., it will continue moving along its current trajectory unless it collides with a paddle or a wall). An *observation function* is a process that transforms latent states of the world to sensory observations. To keep with the Pong example, the latent state of the world contains the true positions of the paddles and the ball, while the observation is the rendered image on the screen.

More formally, the definition stated above can be written as follows (Figure 3.1):

$$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t), \tag{3.1}$$

$$o_t \sim p(o_t \mid s_t), \tag{3.2}$$

[1] I emphasize that this a algorithmic-level definition because there are empirical phenomena associated with mental simulation which do not have clear computational- or algorithmic-level explanations. For example, the effects of gustatory imagery on salivation (K. D. White, 1978) are likely explained best at the implementation level because they are likely side-effects of recruiting the same neural circuitry for both perception and imagery.

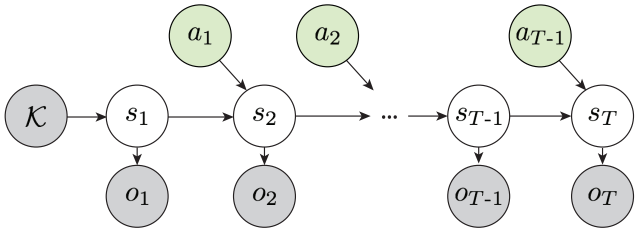


Figure 3.1: Graphical model of mental simulation. This figure shows a number of steps of mental simulation both in terms of the transition function definition (Equations 3.1 and 3.3) and the observation function definition (Equation 3.2). Here, s_t are states, o_t are observations, a_t are actions, and $\mathcal{K}$ is prior knowledge.

where the s_t are states of the world; a_t is an action that affects the world; and o_t is an observation.[2] Equation 3.1 corresponds to the definition of mental simulation as a transition function, and Equation 3.2 to that of mental simulation as an observation function.

Before diving deeper into how Equations 3.1 and 3.2 relate to the empirical phenomena from Chapter 2, I would like to point out two subtleties. First, there is a special case of Equation 3.1 in which there is no previous state but where there is other information or knowledge available (such as the oral description of a scene). I denote this scenario as:

$$s_1 \sim p(s_1 \mid \mathcal{K}), \tag{3.3}$$

where $\mathcal{K}$ stands for "knowledge", but emphasize that this special case still falls under the definition of "transition function". Second, Equation 3.1 also depends on an action, a_t. The choice of action is crucially important when using mental simulation for problem solving and action, but does not constitute simulation itself: the part which corresponds to mental simulation is the *prediction* of what effect that action will have, not the *choice* of action.

An additional point I would like to emphasize is that there is an important difference between the true transition function (which generates experiences we can sense in the real world) and our approximation of the transition function. For example, in the case of "earworms", we do not literally have a full, detailed simulation of musical instruments playing in our heads. However, we do have some approximation of the generative process which may vary in quality depending

[2] Although I use the subscript t here, note that it does not imply that the process of mental simulation must simulate forward in *time*. It is equally valid for the "steps" to be steps in an abstract sense (such as steps of computation), they do not necessarily need to be "timesteps."

on how much experience we have with a particular type of instrument or musical style.

3.1.1 Reinterpreting the empirical phenomena

In the following paragraphs, I detail how the empirical phenomena described in Chapter 2 can be interpreted as mental simulation either via the use of a transition function or via the use of an observation function. Note that I do not intend to suggest that *all* empirical phenomena in a particular domain can be explained via mental simulation. For example, in language comprehension, there are clearly other processes at play such as syntactic parsing. Or, in the case of intuitive physical reasoning, it does seem as if people sometimes rely visual cues rather than simulation (Schwartz & Black, 1999). Rather, my aim here is rather to take those phenomena which *do* seem to be simulation-like, and to interpret them according to the above definitions.

3.1.1.1 Mental imagery

The *phenomenology* of mental imagery can be best explained as using an observation function (Equation 3.2) to simulate a coarse approximation of the perceptual experience generated by some state of the world. The use of the observation function is important here because it allows both a direct comparison between synthetic observations and true observations, as well as the reuse of other perceptual algorithms on the synthetic observations. For example, people can use mental imagery to combine simple shapes into recognizable objects, such as placing a D on top of a J to form an umbrella (Finke & Slayton, 1988). By using the observation function to render the positions and orientations of the shapes into an image, we can then apply some sort of object recognition to the synthetic observation in order to "see" the umbrella.

The actual *dynamics* of mental imagery corresponds to interpretation of mental simulation as a transition function (Equation 3.1). For example, in the case of mental rotation (Shepard & Metzler, 1971), the transition function governs the way in which we expect the mentally rotated object to change as a function of the rotation (corresponding to the action, a) and our visual experience of those changes is again given by the observation function (Equation 3.2). In this particular case, the initial state may be inferred from observations by inverting Equation 3.2.

3.1.1.2 Memory and imagination

As described above, mental simulation is related to episodic memory and imagination in that both of these phenomena involve retrieving pieces of information, stitching them together, and then filling in the missing details (Schacter et al., 2012). This process is analogous to constructing a mental model of an event—or, in other words, constructing a representation of a state of the world conditioned on existing knowledge (Equation 3.3) as well as on the constraints imposed by the transition function itself (Equation 3.1).

3.1.1.3 *Language*

Similar to the processes engaged in memory and imagination, language also induces us to construct a mental model of what is being described (e.g. Zwaan et al., 2002). Language, then, can also be analyzed under the transition function interpretation: it involves a mixture of constructing the initial state (Equation 3.3) and then transitioning to new states as new information is gained (Equation 3.1). For example, upon hearing "they were in a room", the initial state might include some representation of a room with at least two people in it. Upon hearing "the man stood up", the original representation is modified to include exactly one man (sitting down) and at least one woman and/or child; the state then transitions to a new state in which the man is standing.

3.1.1.4 *Intuitive physics*

As described in detail by Battaglia et al. (2013), simulation in physical reasoning corresponds most closely to the interpretation of simulation as a transition function (Equation 3.1), where the laws of physics govern transitions between states. This transition function can then be used to make predictions about the future state of the scene (Battaglia et al., 2013), to reason about what could have happened (Gerstenberg et al., 2014), or to infer unobserved properties of the scene (Chapter 4). The interpretation of simulation as an observation function also plays an important role. Inverting Equation 3.2 enables us to infer the underlying state of the world (subject to physical constraints, such as that objects cannot pass through each other) given an initial observation of the scene.

3.1.1.5 *Intuitive psychology*

Intuitive psychology also can be analyzed using the definition of simulation as a transition function (Equation 3.1). This interpretation is agnostic to whether the transition function is a "theory" of other agents (Gopnik & Wellman, 1992), or whether it is one's own decision-making apparatus (Goldman, 1992; Gordon, 1992): either way, the transition function governs how another agent should act and how its beliefs should change in response to observations, actions, and previous states of the world. This transition function can then be used—similarly to how it is used in intuitive physics—to predict how other agents will behave, to predict what they will think, and to infer what they desire or believe.

3.1.1.6 *Problem solving, thought experiments and counterfactual reasoning*

The areas of problem solving, thought experiments, and counterfactual reasoning often rely on multiple other aspects of cognition including language, intuitive physics, and intuitive psychology. Thus, the sense in which problem solving, thought experiments, and counterfactual reasoning are mental simulations is the same as that for language, intuitive physics, and intuitive psychology. Specifically, the transition function again plays are large role here, governing how the world should change when solving a problem, engaging in a thought experiment, or reasoning about an alternate state of the world.

One "simulation-like" behavior that also falls under this category is that of the sequential application of rules: for example, computing 3×5 by counting up by three, five times. Although more abstract than many of the simulations we have been talking about, this still falls under the notion of a transition function as well. The initial state is the value of the initial counter (zero) and the actions are increments to the counter (add three).

3.1.2 Previous models of mental simulation

For the most part, mental simulation has escaped the attention of cognitive modelers. While there do exist a handful of early models of mental simulation (Funt, 1983; Just & Carpenter, 1976, 1985; Kosslyn & Shwartz, 1977), the majority of existing formal models of mental simulation have been developed within the last decade. While there are distinct differences between both early and recent models of mental simulation, they all tend to adhere to the interpretation of mental simulation as a transition function or an observation function, even if not explicitly stated at such.

The earlier models of mental simulation tended to focus exclusively on the algorithmic level of analysis (Marr, 1982) and did not always make a distinction between states and observations. For example, Kosslyn and Shwartz (1977) describe a model of mental imagery which represents the state directly as an image, which is then manipulated in various ways (e.g. by scaling, translating, or rotating), rather than separating the simulation into separate states (which are manipulated) and then observed (as an image). Funt (1983) takes a slightly different parallel-processing approach to modeling mental rotation, operating on the state using a representation similar to voxels. The model of mental rotation by Just and Carpenter (1976, 1985) comes the closest to explicitly acknowledging states and observations by distinguishing between processes for "search" (finding the same part in each object), "transformation" (determining the transformation required to bring the part in one image into alignment with the other image), and "confirmation" (determining whether the rest of the images are also aligned after the transformation). Here, the search and confirmation steps correspond more to computations taking place over observations, and the transformation step corresponds more to computations taking place over states.

More recent models of mental simulation have made the difference between states and observations explicit. For example, Battaglia et al. (2013) and Smith and Vul (2013) model mental simulation in the domain of physical reasoning using physical simulation as the transition function and an observation function based on the notion of inverse graphics. In other domains, where the perceptual component is less crucial, models exist only of the transition function (though observations may still play a role). For example, Baker, Saxe, and Tenenbaum (2009); Baker and Tenenbaum (2014) model theory of mind using a transition model of how an observed agents' belief changes as a result of its actions and observations. Other models forgo the observation (perceptual) component entirely and rely only on a state representation that encodes abstract spatial information such as position or order (e.g. Khemlani et al., 2013).

3.1.3 What isn't mental simulation?

At this point, the astute reader may be wondering what *isn't* mental simulation. As I have defined it so far, mental simulation seems to encompass every aspect of cognition. Yet, there are many other cognitive phenomena which do not fit neatly into the above definition of mental simulation, such as heuristics or reflexive behaviors. The core component of these other phenomena that seems to be missing—the thing which makes them not mental simulation—is having a causal, generative model of *change*. As discussed previously, this model governs either how the state ought to change over time or as a function of actions (i.e., Equation 3.1), or how the world gives rise to our perceptual observations (i.e., Equation 3.2).

Heuristics clearly fall outside the umbrella of mental simulation as they do not involve a model: for example, in determining how far the blocks of a block tower will scatter when the tower falls, people do not rely on a model of the dynamics of the system, but instead use the heuristic that the blocks will fall further if the tower is taller (Battaglia et al., 2013). This heuristic neither predicts one state from another, nor does it predict sensory observations from an underlying state; rather, the heuristic directly predicts a desired quantity from the current state. More subtly, other forms of causal reasoning which do involve models may fall outside the realm of mental simulation if those models do not govern change. For example, generalization and similarity in pattern matching involves a model of the types of patterns that are likely to occur in the world, but this model says nothing about how the patterns change over time or as a result of actions (Tenenbaum & Griffiths, 2001). In such cases, there are no states to transition between, and no sensory observations generated from underlying states.

The definition of mental simulation as involving a model of change, even if it does exclude some cognitive phenomena, is still so broad that it provides little insight into the specifics of individual mental simulation phenomena. However, I would like to argue that the point of my definition of mental simulation is not to provide new insights on specific phenomena, but rather to provide a new perspective on mental simulation as a whole. By framing mental simulation in terms of transition functions and observation functions, it becomes much clearer what questions have been asked and which have not. In particular, this framing points most obviously towards the meta-level questions surrounding mental simulation, which I discuss further in the next section.

3.2 Metareasoning and Mental Simulation

In the previous section, I defined mental simulation as being either a transition function or an observation function. However, this definition brings to the surface many questions regarding how mental simulation is used, beyond just what mental simulation "is". In this section, I first discuss what some of these unresolved questions are, regarding the representation, control, and use of mental simulations. I then suggest that we can begin to answer some of these questions by focusing on the notion of *metareasoning*, which is how the mind manages its computational resources.

3.2.1 Unresolved questions regarding mental simulation

3.2.1.1 What is the representation for mental simulation?

First, the nature of the simulations themselves are far from being well-understood. While progress has been made in understanding certain types of mental simulations, such as physical reasoning (Battaglia et al., 2013), there are aspects of these simulations that have not yet been explained. For example, why do people's simulations tend toward the center of the screen, as reported by Smith and Vul (2013)? And, how should we computationally model the mental models that people build of others' behavior and preferences? While research suggest ways to model certain aspects of these mental models (e.g. Baker et al., 2009; Baker & Tenenbaum, 2014), we are far from understanding how to construct complete models of individuals.

3.2.1.2 *How do people choose the actions for their simulations?*

Second, how people control their simulations (i.e., how they choose the action, a, in Equation 3.1) is an open question. The answer according to rational analysis (Anderson, 1990) is to formulate this sequential decison-making problem as a Markov Decision Problem (MDP) and solve for the optimal policy (Sutton & Barto, 1998). However, in practice, solving for the optimal policy may be intractable, and this is an open problem actively being researched in computer science. Additionally, formulating the problem as an MDP requires defining a scalar reward function that needs to be maximized. While there are some proposals from the AI literature for how this might be accomplished (e.g. Denil et al., 2017), it is far from clear whether this is actually the "right" solution, or whether it corresponds to how people might choose their reward functions.

One might object that the choice of actions in mental simulation is not fundamentally different from the choice of actions in the real world; thus, it does not truly matter whether we are talking about simulation or not. I would argue, however, that while the choice of actions in mental simulation is similar to the choice of actions in the real world, the two are not the same. For example, although it might be dangerous to take certain actions in the real world (e.g., climbing a cliff face), it is not dangerous to imagine taking those actions. This distinction reveals that the MDPs implied by the real world are different than those implied by mental simulation, and so understanding how people choose actions in the real world does not necessarily imply an understanding of how they choose actions in mental simulation (though I would expect the strategies to be similar).

3.2.1.3 *How do people use mental simulation?*

Third, performing inference over Equations 3.1, 3.2, and 3.3 is often non-trivial and in some cases even intractable (just like solving the relevant MDP, as discussed above). Given the theoretical difficulty of solving these problems, we can safely assume that people are not solving them exactly. One alternative, known as *resource-rational analysis*, posits that people instead compute the solutions to difficult problems using approximate methods, and that the degree of approximation is dependent on how many computational resources are available (Griffiths, Lieder, & Goodman, 2015). This leads to a number of additional questions: what should the representation be that the simulation operates over, which approximate inference algorithm should be used, and how should the hyperparameters of that algorithm be set? For example, when reasoning about a tower of building blocks such as those studied by Battaglia et al. (2013), should the towers be represented as a collection of 10 individual blocks, or are blocks that behave similarly grouped together into a sort of "megablock"? Should the representation even be object-based at all? How many simulations should be run of the dynamics of the tower, and at what resolution?

3.2.2 Metareasoning

When we talk about questions concerning the choice of representation, the control of an algorithm, or the tuning of hyperparameters to solve a problem, we are really referring to:

***Metareasoning*: how the mind manages its computational resources.**

Metareasoning is a term from the AI literature (Russell & Wefald, 1991) that has recently been adopted in cognitive science (Gershman, Horvitz, & Tenenbaum, 2015; Lieder et al., 2014). In the context of this thesis, we can think about mental simulation as one particular type of computational resource. Thus, in this case, metareasoning corresponds to the meta-level decisions that need to be made about *how* to use mental simulation: what representation should be chosen, how long the simulation should be run for, how many simulations should be run, and which simulations should be chosen? To fully understand mental simulation, it is not enough to demonstrate that it used: we must also understand the manner in which it is used, and how people make the relevant meta-level decisions about how to use it in those ways.

In the remainder of this thesis, I focus on how we can begin to answer the meta-level questions I have posed in this section. My hope is that the research presented in this thesis will begin to move the conversation away from the specifics of what mental simulation *is*, to *how* it is used. Specifically, I focus my efforts on three questions in particular: how people use simulation for tasks other than prediction (Chapter 4); how people choose what simulations to run (Chapter 5); and how people decide how many simulations to run (Chapter 6). Finally, I use the results from investigating these questions to help develop a more detailed theory of metareasoning with respect to mental simulation (Chapter 7).

Imagination is the Discovering Faculty, pre-eminently.

Ada Lovelace

4 Learning by Thinking

Consider the scene in Figure 4.1a. Despite the difference in size, one can infer that the mass of the forklift is large compared to that of the storage container. Inferences about the physical properties of objects such as mass and friction are critical to how we understand and interact with our surroundings. While they are sometimes specified unambiguously by a small set of perceptible features such as size, material, or tactile sensations, we often access them only indirectly via their physical influence on observable objects. Here, we ask: how do people make such inferences about the unobservable physical attributes of objects from complex scenes and events?

In addition to one-off inferences about properties such as mass, people form beliefs about these properties over time. For example, through experience, people learn that certain materials (e.g., metal) are heavier than others (e.g., plastic). How is it that people learn these attributes? Certainly, people may rely on sensorimotor feedback as they hold and manipulate objects (e.g. Baugh, Kao, Johansson, & Flanagan, 2012). Can people also learn through experience if only visual information about the static and dynamic behavior of such objects is available? If so, what is the mechanism by which they do this?

There is a vast literature on whether (and if so, how) people reason about mass. People are

The text of this chapter was previously published as Hamrick, Battaglia, Griffiths, and Tenenbaum (2016). Credit goes to Lombrozo (forthcoming) for the term “learning by thinking”.

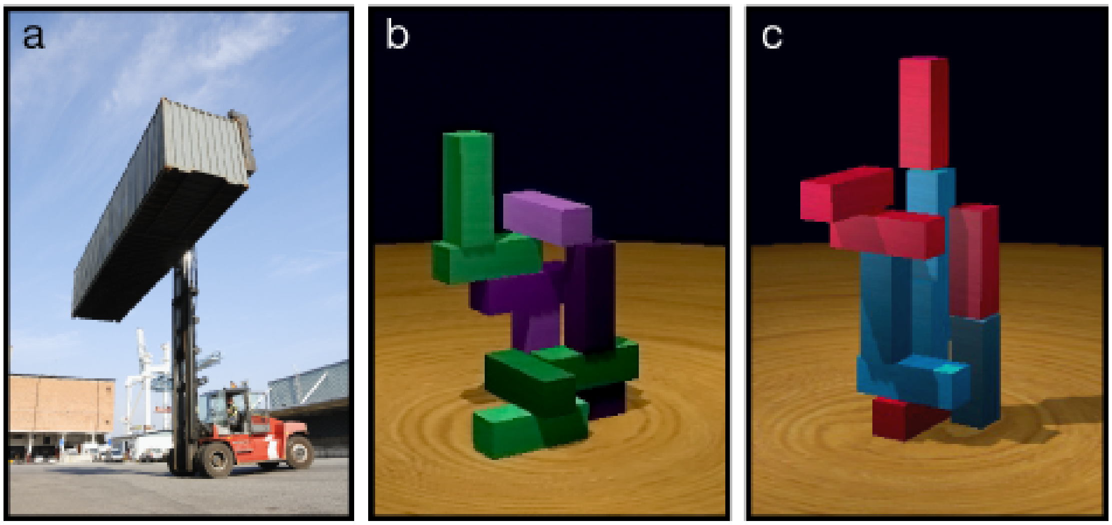


Figure 4.1: Three scenes that engage our ability to reason about mass. (a) The forklift's weight counterbalances the container's. (b-c) Two examples of experimental stimuli. If the green blocks in (b) are heavier than the purple blocks, you can predict that the tower will fall down rather than remain standing. If the tower in (c) stays standing, you can infer that the blue blocks are heavier.

clearly sensitive to mass when reasoning about other physical properties: for example, people's memory for the location of an object is affected by its implied weight (Hubbard, 1997); similarly, people make different judgments about how a tower of blocks will fall down depending on which blocks they think are heavier (Battaglia et al., 2013). Previous studies of how humans *infer* mass from observed collision dynamics have examined the relative roles of perceptual invariants (Runeson et al., 2000) and heuristics (Gilden & Proffitt, 1994; Todd & Warren, 1982), focusing on judgments about simple one- or two-dimensional (1D or 2D) situations with one or two objects. However, the real world is much more complex: everyday scenes are three-dimensional (3D) and often involve many objects.[1] Moreover, collisions between objects are not the only factor affecting peoples' judgments: for example, there are no collisions in the forklift scene in Figure 4.1a, yet we can easily infer what the relative masses of the objects might be.

A question related to *whether* people can make accurate inferences about unobservable physical properties is *how* they make any inferences at all. Sanborn et al. (2009; 2013) proposed that inferences could be characterized by a model that performs Bayesian inference over structured

[1] We define a 3D scene to be any scene that contains depth information, regardless of whether it is viewed as a 2D projection. We define a 2D scene to be a scene with no depth cues (i.e., it is truly 2D and not a projection of a 3D scene).

knowledge of Newtonian physics[2] and noisy or uncertain perceptual inputs. In the 2D case, this "Noisy Newton" hypothesis works well for inferring properties like mass because the laws of Newtonian physics (such as conservation of momentum) can be encoded as distributions over random variables such as velocity, where the randomness comes from perceptual uncertainty (Sanborn, 2014; Sanborn et al., 2009, 2013). However, for scenes involving both statics and dynamics, it is not clear where these probabilities should come from. For example, if the forklift in Figure 4.1a is about to tip over, you can infer that the storage container is heavier, because if it were not, the forklift would likely remain upright. Where does this "likelihood of remaining upright" come from?

Recent research has proposed that people reason about complex environments using approximate and probabilistic mental simulations of physical dynamics (Battaglia et al., 2013; Hamrick, Battaglia, & Tenenbaum, 2011). They are *approximate* in the sense that they do not analytically solve the exact equations that underly Newtonian physics, but rather estimate the implications of those equations through an iterative process. They are *probabilistic* in that the simulations are non-deterministic, where the stochasticity reflects uncertainty that arises from noisy perceptual processes and imperfect knowledge of the scene. There is a growing body of evidence that people use such approximate and probabilistic mental simulations, including explanations of human judgments of physical causality and prediction in a wide range of scenarios (Gerstenberg et al., 2012, 2014; Smith, Battaglia, & Vul, 2013; Smith, Dechter, et al., 2013; Smith & Vul, 2013; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2014). A similar hypothesis has also been proposed by P. A. White (2012), suggesting that simulations are the result of retrieving past perceptual experiences and extrapolating using a forward model.

If people use probabilistic mental simulations to make predictions about physical scenes, then it should be possible for people to use those simulations to estimate the probabilities of different outcomes. These probabilities can be used to make inferences about unobservable physical properties. Indeed, recent work by Ullman et al. (2014) and Gerstenberg et al. (2012, 2014) has provided examples of how simulations might be used in simple 2D scenes to estimate the necessary probabilities for Bayesian inference. However, such an approach has not been applied to the types of complex, 3D scenes that people encounter in the real world.

Using probabilistic simulation to make inferences about unobservable physical properties also suggests a unified framework both for reasoning about individual object-level properties (i.e.,

[2] In this context, we take "structured" to mean implicit knowledge of formal physical laws, in contrast to implicit knowledge of naïve physics or explicit knowledge of formal physics. See the General Discussion for further discussion of how these differing forms of physical knowledge relate.

that the forklift in Figure 4.1a is heavier than the storage container) as well as class- or material-level properties (i.e., that objects made out of stone are heavier than objects made out of plastic). Historically, research has focused on how people make one-shot inferences about the properties of individual objects (Gilden & Proffitt, 1989a; Runeson et al., 2000; Sanborn, 2014; Sanborn et al., 2009, 2013; Todd & Warren, 1982), but not on how these inferences might also play a role in learning class-level properties such as the density of a particular material. We suggest that if Bayesian inference is performed using probabilities obtained through approximate physical simulation, then this could provide an account for *both* one-shot inferences and learning. Specifically, Bayes' rule dictates both how to compute inferences about individual objects, as well as how to integrate multiple pieces of information over time to learn about the properties of classes of objects.

This work is the first to explore people's ability to make inferences about mass in complex scenes that may be either static or dynamic, and addresses two questions regarding this ability. First: *can* people make accurate inferences? To answer this, we present three experiments in which we asked people to make inferences about the relative masses of objects in complex scenes involving both static and dynamic objects. We find that people can form accurate judgments about the relative mass, and that they become increasingly fine-tuned to these properties as they accumulate multiple pieces of information. Second: *how* do people make inferences about properties like mass? We introduce a new cognitive model that uses approximate, probabilistic simulation to estimate probabilities needed by Bayesian inference to produce judgments about the relative mass of objects. When compared to data from our experiments, we find that our model is a good characterization of how people make inferences about the masses of individual objects and how they learn about the mass of a class of objects. Moreover, by replacing the model's simulations with people's own predictions about the future dynamics of the scenes, our model can predict people's inferences about mass with high accuracy. This suggests that the same simulation-based mechanism the mind uses for predicting physics is also involved in forming physical inferences about object-level properties and in learning the properties of a class of objects over time.

4.1 Experiment 4.1: Inferring Mass From a Single Trial

We first asked the question: *can* people infer mass in complex scenes? To answer this, we ran three experiments in which we showed people videos of towers of blocks, where each block was one of two colors, and asked them to judge which color was heavier. This section describes the

Table 4.1: Phases of Experiment 4.1.

	Phase				
	Familiarization	*Training*	*Prediction*	*Inference*	*Posttest*
Stimuli	control	training	experimental	experimental	control
Block color	random	red/blue	red/blue	varied	random
Judgment	fall?	fall?	fall?	heavier?	fall?
Feedback	yes	yes	no	-	yes

first of these experiments.

4.1.1 Participants

We recruited participants on Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015). We randomly assigned participants to one of four conditions, which determined which stimuli they saw (see Design for details). Based on sample sizes from earlier pilot studies, we aimed to have 20 participants per condition that would result in 40 judgments per stimulus. However, also based on earlier pilot data, we expected about 20% of participants to fail an attention check (see the *posttest* in the Procedure section for details). Thus, we aimed for 100 participants overall, though we ended up recruiting 101 participants because one participant had completed an earlier version of the experiment. We excluded this one participant from analysis, as well as 19 participants for failing the attention check. All participants were treated in accordance with UC Berkeley's IRB protocols and were paid $1.25. All participants were 18 years or older, and completed the experiment from within the United States. Participants took a median of 14.6 minutes to perform the full experiment.

4.1.2 Stimuli

Stimuli were videos of computer generated 3D towers of identically sized building blocks arranged in a way such that both the position and orientation of the blocks were relevant to the dynamics of the scene (see Appendix A.1.1 for details on how stimuli were constructed, and see Figure 4.1b-c for two example stimuli). Stimulus presentation videos showed the camera (beginning from a random angle) rotating 180° counterclockwise around the tower for 5 seconds, during which gravity was set to zero so that none of the towers would fall down. Separate feedback videos showed the physical dynamics of the tower falling or not falling for 3.5 seconds under a gravitational acceleration of $-9.81\frac{m}{s^2}$. The feedback videos began from the last frame of the stimulus presentation video that depicted the same tower.

4.1.2.1 *Experimental stimuli*

There were 20 experimental towers, which consisted of 5 blocks of one color and 5 blocks of another color (see Figure 4.1b-c for two examples). The relative mass ratio between the differently-colored blocks was either 1:10 or 10:1, which is approximately the same as that between a metal steel block and a wooden oak block. We chose the experimental towers such that whether the tower fell or not was determined both by the geometry of the structure, as well as the mass ratio.[3] For example, if a tower had a mass ratio of 1:10 and fell down, then a tower with the same geometry but a mass ratio of 10:1 would stay standing (see Appendix A.1.2 for details). Note that this has the consequence that for all the towers we chose, if the tower fell under one mass ratio, it would *not* fall under the other, and vice versa (see the section on "Reasoning About Physical Properties via Simulation" for further discussion of the consequence of this choice).

4.1.2.2 *Training stimuli*

There were also ten training stimuli, which were of the same form as the experimental stimuli (5 blocks of one color and 5 blocks of another color, with relative masses of 1:10 or 10:1).

4.1.2.3 *Control stimuli*

Finally, there were six control stimuli, which consisted of 10 randomly-colored blocks that all had the same mass and which were taken from a previous experiment (Battaglia et al., 2013). We chose the control towers to be those towers which participants rated in that experiment as extremely stable or extremely unstable.

4.1.3 Design

The experiment consisted of five phases: *familiarization*, *training*, *prediction*, *inference*, and *posttest*. Within each phase, the trial order was randomized for each participant. Participants were assigned randomly to one of four conditions, which determined the mass ratios of the towers in the training, prediction and inference phases. Specifically, we constructed a 2×2 design

[3] The stability of a given tower is strongly dependent on perceptual information, such as its geometry, as well as explicit information, such as which blocks participants think are heavier. See Battaglia et al. (2013) for a detailed investigation into what determines the stability of a tower.

in which participants observed towers with a mass ratio of either 1:10 or 10:1 in the training and prediction phases, and towers with a separate mass ratio of either 1:10 or 10:1 in the inference phase. A summary of each phase is provided in Table 4.1.

4.1.3.1 Familiarization phase

The familiarization phase familiarized participants with the "will it fall?" decision. They answered this question for the six control stimuli, and received feedback after responding.

4.1.3.2 Training phase

The training phase familiarized participants with the task of judging "will it fall?" when the blocks could have different masses. Participants answered "will it fall?" for the 10 training stimuli, which had blocks colored red and blue; participants were told which color block was heavier. Depending on condition, the mass ratio between the blocks was either 1:10 or 10:1, and the colors were counterbalanced. Participants received feedback after responding.

4.1.3.3 Prediction phase

In the prediction phase, participants again judged "will it fall?", but for the 20 experimental stimuli. This phase was included in order to *a priori* estimate how likely people thought it was for towers to fall under different mass ratios. The blocks in these stimuli were also colored red and blue, and had the same mass ratio as in the training phase. Participants did not receive any feedback.

4.1.3.4 Inference phase

The inference phase was designed to gather participants' inferences about the relative mass ratios of the blocks. Participants answered "which is the heavy color?" for the same 20 experimental stimuli (in a different order) as in the prediction phase after observing a video of the tower falling or not falling. The colors of the blocks changed on every trial, and no pair of colors was shown more than once, though individual colors were reused in different pairs of colors (see Appendix A.1.3 for all pairs of colors). Depending on condition, the mass ratio between the blocks was either 1:10 or 10:1, and was either the same or different as the ratio in the training and prediction phases. As before, whether a given color was assigned to be heavy or not was counterbalanced across participants.

4.1.3.5 *Posttest phase*

The posttest phase was identical to the familiarization phase, with the exception of trial order. The purpose of the posttest was to check that participants were paying attention by asking them to perform more stability judgments. This design was motivated by the assumption that by the end of the experiment, participants should be able to judge the easier training towers with high accuracy. We excluded participants from analysis who incorrectly judged the stability of at least one (out of six) towers in the posttest.

4.1.4 Procedure

There were two types of trials: *stability* trials, in which participants were asked to predict whether the towers would fall down, and *mass* trials, in which participants were asked to infer which color block was heavier. Participants initiated all trials by pressing the 'c' key, after which the stimulus presentation began.

4.1.4.1 *Stability trials*

On stability trials, participants were then asked the question, "On a scale from 1-7, how likely is the tower to fall down?", where 1 meant "unlikely to fall" and 7 meant "likely to fall". Participants responded by pressing the corresponding number key. Participants then saw feedback (if any, depending on the phase) immediately after responding. Feedback consisted of a video depicting the tower either falling or not falling was shown in addition to text indicating "tower falls" or "tower does not fall".

4.1.4.2 *Mass trials*

On mass trials, after being shown the stimulus presentation, participants were prompted to press the 'c' key to view feedback. After the feedback was complete, they were asked the question, "Which is the heavy color?", and could click one of two buttons corresponding to the block colors.

4.1.5 Analysis

Before presenting the results, we describe a few generic analyses that we perform several times in the subsequent results section. For analyses of participant's accuracy, we computed medians and 95% confidence intervals using 10000 bootstrap samples. To test if people's judgments were above chance on a particular stimulus, we used the same bootstrap analysis and tested whether

$p\left(p(\text{correct}) \leq 0.5\right) \leq \frac{0.05}{40}$, where $p(\text{correct})$ is an empirical probability of answering correctly and where $\frac{1}{40}$ is a Bonferroni correction for multiple comparisons. That is, this equation is a test for whether the empirical frequency of answering correctly is significantly greater than 0.5 (chance), where "significantly" is defined as $p < 0.05$, adjusted for multiple comparisons. For correlation analyses, we used a bootstrap analysis of 10000 bootstrap samples to compute the median and 95% confidence intervals of both Spearman (ρ) and Pearson (r) correlations. Any other reported confidence intervals were similarly computed using 10000 bootstrap samples with replacement.

4.1.6 Results

Overall accuracy in responding to "which is the heavy color?" was significantly above chance ($M = 81.9\%$, 95% CI $[79.9\%, 83.7\%]$, averaged across participants and stimuli), though there were 9 individual stimuli (out of 40) for which accuracy was not significantly above chance (corrected for multiple comparisons). We suspect that the stimuli which participants did not classify above chance were not classified as such either due to a combination of (1) insufficient power and (2) the information in the stimulus presentation genuinely not being very informative. This latter possibility is supported by our cognitive model later in the text (see the section on "Reasoning About Physical Properties via Simulation" and Figure 4.4). Accuracy of individual participants ranged from 45% to 100%, and 95% of participants answered correctly on at least 60% of the trials.

We also looked at how well participants' responses predicted each other, both in response to "will it fall?" and "which is the heavy color?". To compute this, we performed a bootstrap analysis in which each bootstrap sample was computed by randomly dividing the participants in half and calculating the correlation of average judgments between the two groups. For "will it fall?" judgments, this split-half correlation was $r = 0.90$, 95% CI $[0.84, 0.94]$, implying that people were very consistent with each other. Similarly, for "which is the heavy color?" judgments (where 0 corresponded to a ratio of 1:10 and 1 corresponded to a ratio of 10:1), the correlation was $r = 0.95$, 95% CI $[0.92, 0.97]$, again indicating that people were highly consistent. For accuracy in judging the mass[4] (where 0 corresponded to an incorrect answer, and 1 to a correct answer), people were more variable, with a correlation of $r = 0.69$, 95% CI $[0.55, 0.80]$.

To check whether participants improved over time due to practice effects, we computed the Spearman rank correlation between trial number and mean accuracy. We found no significant

[4] Note that here, sometimes the correct ratio was 1:10 and sometimes it was 10:1. Therefore average judgments to "which is the heavy color?" are not the same as the average *accuracy* of those judgments.

effect of practice ($\rho = 0.07$, 95% CI $[-0.20, 0.34]$).

4.1.7 Discussion

In this experiment, participants inferred which blocks were heavier with high accuracy even though they were only shown one example tower per judgment. We next asked: can people accumulate information over time and further improve their inferences based on multiple examples? To answer this, we ran another experiment similar to Experiment 4.1. The main difference was that during the inference phase, the block colors did not change, and participants were told that the mass ratio remained the same.

4.2 Experiment 4.2: Learning From Multiple Trials (Within Subjects)

4.2.1 Participants

As in Experiment 4.1, we recruited participants from Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015) and assigned participants randomly to one of four conditions. We aimed to have 20 participants per condition, and expected 20% of participants to fail the attention check. Thus, we set our target sample size to 100 participants. Overall, we collected data from 111 participants because 7 participants had incomplete or duplicate data due to an experimental error and 4 participants had already done an earlier version of the experiment. We excluded these participants from analysis, as well as 15 participants for failing the attention check. Participants were treated in accordance with UC Berkeley's IRB protocols and were paid $1.00. All participants were 18 years or older and completed the experiment from within the United States. Participants took a median of 16.2 minutes to perform the full experiment.

4.2.2 Design

The design for Experiment 4.2 was identical to that of Experiment 4.1, except during the inference phase. As in Experiment 4.1, we did not tell participants which color was heavier, but we did tell them that the heavier blocks would always have the same color. Instead of having different color pairs on every trial, the colors were always purple and green (e.g., Figure 4.1b), and these colors were counterbalanced across conditions. Participants judged the relative mass only on trials 1, 2,

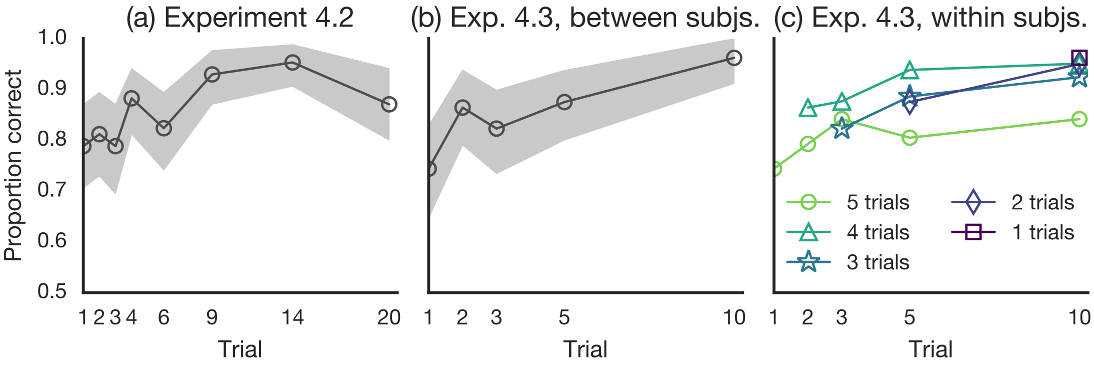


Figure 4.2: Inferences of relative mass in Experiments 4.2-4.3 as a function of trial. In all plots, the solid lines indicate the mean proportion of correct human responses to ``which is the heavy color?''. Shaded regions are 95% confidence intervals of the mean. (a) The left subplot shows accuracy as a function of trial in Experiment 4.2. (b) The middle subplot shows between-subjects accuracy in Experiment 4.3. (c) The right subplot shows within-subjects accuracy in Experiment 4.3.

3, 4, 6, 9, 14, and 20. These trials were chosen based on the hypothesis that participants' beliefs would change a lot at the beginning of the experiment and less so at the end of the experiment; thus, it would be better to ask more frequently at the beginning of the experiment rather than after a fixed interval. On trials where they were not asked, they just watched the stimulus presentation and feedback video, and then immediately moved on to the next trial.

4.2.3 Results

As in Experiment 4.1, participants were above chance in judging which color was heavier overall ($M = 85.6\%$, 95% CI $[82.9\%, 88.2\%]$, averaged across participants and stimuli), though there were 14 individual stimuli for which this was not significant (corrected for multiple comparisons, see the Analysis section of Experiment 4.1). Accuracy of individual participants ranged from 25% to 100%, and 95% of participants answered correctly on at least 50% of the trials.

Unlike Experiment 4.1, participants in Experiment 4.2 were told that the mass of the blocks remained the same during the inference phase. Thus, we would predict their accuracy to generally increase as a function of trial. This trend does appear ($\rho = 0.68$, 95% CI $[0.25, 0.90]$), though as shown in Figure 4.2a, is not monotonic.

Participants were very self-consistent in their responses to "will it fall?", with a split-half correlation of $r = 0.90$, 95% CI $[0.84, 0.94]$. Their responses were also consistent with those in

Experiment 4.1, with a correlation of $r = 0.92$, 95% CI $[0.86, 0.96]$.

4.2.4 Discussion

Why did the accuracy in judging the mass ratio go *down* between trials 14 and 20? If participants were learning over time, we would expect the last trial to have the highest accuracy. We suspected that participants may have been confused by being asked "which is the heavy color?" multiple times, perhaps thinking that they needed to change their response. Alternatively, it is possible that the delay between responding to "which is the heavy color?" might have resulted in a decay of any learning that did happen. To address these issues, we ran a third experiment similar to Experiment 4.2 that was shorter and in which we varied the number of times participants were asked "which is the heavy color?".

4.3 Experiment 4.3: Learning From Multiple Trials (Between Subjects)

4.3.1 Participants

We recruited participants on Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015). Participants were randomly assigned to one of ten conditions (see Design for details), and to keep sample sizes consistent with the first two experiments, we aimed to have 40 participants per condition. Based on Experiments 4.1-4.2, we expected that about 17% of participants would fail the attention check. Thus, we aimed for 480 participants total, though we actually collected data from 487 participants because 7 participants had completed an earlier version of the experiment. We excluded these participants from analysis, as well as 79 participants who failed the attention check. Participants were treated in accordance with UC Berkeley's IRB protocols and were paid $0.70. All participants were 18 years or older and completed the experiment from within the United States. Participants took a median of 8.3 minutes to perform the full experiment.

4.3.2 Design

Experiment 4.3 utilized nearly the same design as Experiment 4.2, except that participants did not complete the prediction phase, and they only completed 10 trials of the inference phase (randomly selected from the full set of 20 stimuli). Rather than make judgments on the exact same trials, five different subgroups of participants were asked to judge the mass different numbers of times: on trials 1, 2, 3, 5 and 10; on trials 2, 3, 5 and 10; on trials 3, 5 and 10; on trials 5 and 10; and just on trial 10. Crossed with the two mass ratios (1:10 and 10:1), this led to a total of ten conditions in Experiment 4.3. This design was chosen in order to isolate the effect of asking the question of "which is the heavy color?" multiple times. If participants were getting confused by being asked the question multiple times, then the participants in this experiment who answer the question first on the 3rd trial, for instance, should have a higher accuracy than those that answer first on the 1st or 2nd trials. Additionally, this design tests the alternate hypothesis that learning could be decaying due to the gap between questions: if learning is decaying, then participants who answer only on the 10th trial should arguably do worse than any of the other conditions.

4.3.3 Results

As with the previous experiments, participants were above chance in judging which color was heavier across all stimuli ($M = 87.1\%$, 95% CI $[85.2\%, 88.9\%]$, averaged across participants and stimuli). There were only 2 individual stimuli for which people's judgments were not significantly above chance (corrected for multiple comparisons, see the Analysis section for Experiment 4.1).

To judge the effect of learning over time, we computed participant's accuracy as a function of trial both between and within subjects. To compute the between subjects accuracy, we took only the first responses from each condition (i.e., only the first time each participant answered "which is the heavy color?"). The Spearman rank correlation between subjects was significant ($\rho = 0.70$, 95% CI $[0.40, 1.00]$), while the rank correlation for the condition in which participants responded on five trials was not ($\rho = 0.60$, 95% CI $[-0.31, 1.00]$). Figure 4.2b-c shows both the between- and within-subjects accuracy as a function of trial.

4.3.4 Discussion

These results suggest that demand effects were at play both in Experiment 4.2 and within subjects in Experiment 4.3. However, the structure of Experiment 4.3 allowed us to perform between-subject analyses, revealing that when participants were not biased by being asked the same question multiple times, they did become increasingly accurate over time. The fact that people are able to incorporate this information over time suggests a way in which people might learn about the properties of classes of objects (for example, the densities of particular materials). Additionally, our results rule out the hypothesis that learning decayed in the gaps between when we asked them which color they thought was heavier, because participants who responded on fewer trials (e.g. only trial 10) did better than those that responded on more trials (e.g. trials 1, 2, 3, 5, and 10).

4.4 Reasoning About Physical Properties via Simulation

The three experiments described previously demonstrate that people *can* make accurate inferences about mass in complex scenes, and that they can accumulate evidence over time. Our next question was: *how* do people infer mass? We hypothesized that people's inferences can be characterized using Bayesian inference in which the probabilities are computed via probabilistic simulation. Here, we formalize this hypothesis in a model observer and compare judgments from the model with people's judgments in Experiments 4.1-4.3.

4.4.1 Observer Model

On each trial, the observer model views a stimulus (S) and receives feedback (F). The feedback is a Bernoulli random variable indicating whether the tower fell ($F = 1$) or did not fall ($F = 0$). Let κ be a random variable corresponding to the mass ratio, and let a particular hypothesis about the mass ratio be indicated by $\kappa = k$ (i.e., either $\kappa = 0.1$ or $\kappa = 10$). Then, the probability of that hypothesis given the observed feedback is computed from Bayes' rule:

$$p(\kappa|F,S) = \frac{p(F|S,\kappa=k)p(\kappa=k)}{\sum_{k'} p(F|S,\kappa=k')p(\kappa=k')}, \tag{4.1}$$

where $p(\kappa = k)$ is the prior probability of the hypothesis that $\kappa = k$, and $p(F|S, \kappa = k)$ is the probability of observing the feedback given that the hypothesis $\kappa = k$ is true.

Equation 4.1 demonstrates how the observer model computes its belief about the mass ratio after a single trial. We can further extend this model to show how an observer model should update its beliefs over time as it encounters more evidence. Specifically, the observer model should use the posterior distribution of each trial as the prior distribution of the next:

$$\begin{aligned} p_t(\kappa=k) &= \frac{p(F_t|S_t,\kappa=k)p_{t-1}(\kappa=k)}{\sum_{k'} p(F_t|S_t,\kappa=k')p_{t-1}(\kappa=k')} \\ &= \frac{p_0(\kappa=k)\prod_{i=1}^t p(F_i|S_i,\kappa=k)}{\sum_{k'} p_0(\kappa=k')\prod_{i=1}^t p(F_i|S_i,\kappa=k')}, \end{aligned} \tag{4.2}$$

where $p_t(\kappa = k) := p(\kappa = k|F_1, S_1, \dots, F_t, S_t)$ is the belief about the mass ratio after observing t trials. Thus, $p_0(\kappa = k)$ is the prior, $p_1(\kappa = k)$ is the belief after the first trial, and so on. Note that each $p_t(\kappa = k)$ is defined recursively in terms of $p_{t-1}(\kappa = k)$ and thus contains all information observed so far up through trial t. This model is consistent with the optimal inference computation for an observer that aggregates information across trials.

We can contrast the model defined in Equation 4.2 (henceforth referred to as the *learning* model) with an observer that does not accumulate information over trials. This *static* model does make inferences according to Equation 4.1, but effectively starts from scratch on each trial:

$$p_t(\kappa = k) = \frac{p(F_t|S_t, \kappa = k)p_0(\kappa = k)}{\sum_{k'} p(F_t|S_t, \kappa = k')p_0(\kappa = k')}. \tag{4.3}$$

As above, p_0 is the prior. This model is consistent with the optimal inference computation for an observer that does not aggregate information across trials.

4.4.2 Estimating Probabilities with Simulation

We posit that people compute the likelihood term $p(F_t|S_t, \kappa = k)$ in Equation 4.1 using their "intuitive physics engine" (IPE), as proposed by Battaglia et al. (2013) and Hamrick et al. (2011). We will refer to this version of the observer model as the "IPE observer model". The IPE is a hypothetical cognitive mechanism that makes predictions by running forward noisy physical simulations. Because these simulations are non-deterministic, running many under different values of κ allows people to estimate how likely a particular outcome is under each hypothesis.

The IPE observer model runs N simulations and, for each simulation, compares the initial and final states of the stimulus. Specifically, the estimated probability of observing F_t given a stimulus S_t and the mass ratio κ is:

$$p(F_t|S_t, \kappa = k) \approx \frac{1}{N}\sum_{i=1}^{N} b^{(i)}, \tag{4.4}$$

where $b^{(i)} \sim \mathrm{IPE}(S_t, \kappa)$ is the fraction of blocks (ranging from 0 to 1) that moved more than 0.25 cm from their initial positions during the i^{th} simulation. There are other possible "queries" that could be used to define what it means for a tower to fall, such as whether any blocks moved at all. Such alternate queries yield similar results and are discussed further in Appendix A.3.

4.4.3 Empirically Estimating Probabilities

As stated previously, we hypothesize that people are using an IPE to predict how likely it is for a tower to fall. If this is the case, and if we use probabilities estimated from their predictions of "will it fall?" in our learning and static models, then these empirically-based models should predict people's inferences of mass just as well (or better) than those which use the IPE model's predictions. To test this hypothesis, we computed Equation 4.1 based on "will it fall?" judgments from the prediction phases of Experiments 4.1-4.2; we will refer to this version of the model as the "empirical observer model". In order to turn these judgments into probabilities, we assumed that the smallest response (1) was equivalent to 0 and the largest response (7) was equivalent to 1. By rescaling these judgments and averaging, we obtained an empirical estimate of the probability that the tower will fall, analogous to Equation 4.4:

$$p(F_t|S_t, \kappa = k) \approx \frac{1}{M}\sum_{i=1}^{M}\frac{J_{\mathrm{fall}}^{(i)} - 1}{6}, \tag{4.5}$$

where M is the number of participants and $J_{\mathrm{fall}}^{(i)}$ is the judgment of the i-th participant. This assumes participants are running only one simulation to make their judgments, which is not necessarily the case (Battaglia et al., 2013). While running more simulations would not change the overall mean of the likelihood, it would change the variance, and this possibility is investigated in more depth in the General Discussion.

4.4.4 Fitting Model Parameters

We fit both the static and learning models to human data using a Bayesian logistic regression. More formally, let the i^{th} participant's judgment of the mass ratio on trial t be a Bernoulli random variable denoted by $J_{\mathrm{mass},t}^{(i)}$. Then, assume that a participant's judgment is related to their belief via the logistic function:

$$p(J_{\mathrm{mass},t}^{(i)} = k|\beta, \eta_t) = \frac{1}{1 + e^{-\beta\eta_t}}, \tag{4.6}$$

where β is a parameter that specifies how strongly the evidence is weighed and η_t is the posterior log odds at time t. Specifically, in the case of the learning model, we have:

$$\eta_t = \frac{p(F_t|S_t, \kappa = 10)p_{t-1}(\kappa = 10)}{p(F_t|S_t, \kappa = 0.1)p_{t-1}(\kappa = 0.1)}. \tag{4.7}$$

The equation for the posterior log odds is the same for the static model, except that $p_{t-1} = p_0$ for all t. Given this formulation, we are interested in finding the value of β which best fits a participant's responses, i.e., the maximum *a posteriori* (MAP) estimate over $p(\beta|J^{(i)}_{\text{mass},1}, \ldots, J^{(i)}_{\text{mass},T}, \eta_1, \ldots, \eta_T)$. To avoid overfitting, we used a Laplace prior with $\mu = 1$ and $b = 1$ (equivalent to L1 regularized logistic regression). Then, we found the MAP estimate of β for each participant separately.

In this regression, the β parameter reflects how strongly participants weigh the evidence they have observed. In the case of the static model, this translates to just how strongly they take into account the evidence on each trial; in the learning model, this translates to a learning rate. A value of $\beta = 1$ means that the model weighs evidence in accordance to Bayes' rule. A value of $0 < \beta < 1$ means that the model weighs the evidence *less* strongly than Bayes' rule does, but does still take it into account. A value of $\beta > 1$ means that the model weighs evidence *more* strongly than Bayes' rule does. A value of $\beta = 0$ means that the model ignores all evidence, and a value of $\beta < 0$ means that the model believes the *opposite* of what the evidence tells it.

Although the Laplace prior works to prevent overfitting, we were still concerned about the possibility of overfitting, particularly in Experiment 4.3 where we had only one or two responses from some participants. Thus, to be able to compare our models without making assumptions about which parameter values were correct, we computed Bayes factors (Kass & Raftery, 1995). Briefly, a Bayes factor is defined as the marginal likelihood ratio of the data given two different models, with the parameters of those models integrated out. Thus, the Bayes factor gives a measure of how much better one model explains the data over another model, irrespective of the specific parameter values of the model. According to Kass and Raftery (1995), the value of the log of the Bayes factor can be interpreted as positive evidence if $1 < \log K \leq 3$, strong evidence if $3 < \log K \leq 5$, and very strong evidence if $\log K > 5$.

We computed our Bayes factors in the following manner. We first computed the marginal likelihood of participants judgments under each model by integrating over possible values of β:

$$p(J_{\text{mass}}|\eta) = \prod_{i=1}^{N} \left(\int \prod_{t=1}^{T} p(J^{(i)}_{\text{mass},t} = k|\beta^{(i)}, \eta^{(i)}_t) p(\beta^{(i)}) \, \mathrm{d}\beta^{(i)} \right), \tag{4.8}$$

where N is the number of participants, T is the number of trials, $J^{(i)}_{\text{mass},t}$ is the judgment of the i^{th} participant on trial t, $\eta^{(i)}_t$ is the posterior log odds of participant i on trial t, and $\beta^{(i)}$ is the parameter for participant i. We then computed Bayes factors by computing the log ratio between the marginal likelihoods for the learning model to the static model.

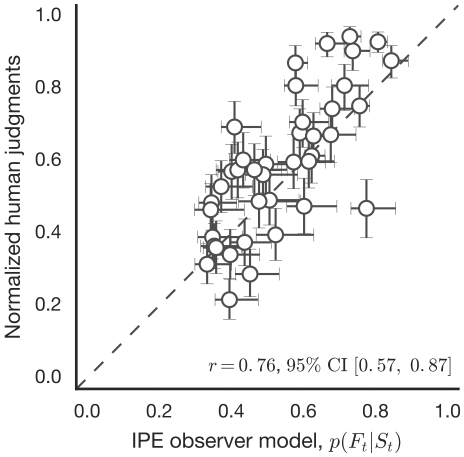


Figure 4.3: Responses to ``will it fall?'' in Experiments 4.1-4.2. The x-axis shows responses from the IPE model observer, and the y-axis shows responses from participants. Error bars are bootstrapped 95% confidence intervals. The dashed line indicates perfect correspondence between the model and people.

4.4.5 Results

4.4.5.1 *Predictions*

We first checked whether people's responses to "will it fall?" in the prediction phase were consistent with previous findings (Battaglia et al., 2013). We pooled responses to "will it fall?" from the prediction phases of Experiments 4.1-4.2, used them to compute Equation 4.5, and then compared them to IPE predictions (computed fation 4.4). We found that participants were well-predicted by the IPE model ($r = 0.75$, 95% CI $[0.57, 0.87]$), and note that this correlation is about the same as that found by Battaglia et al. (2013), which was $r = 0.80$, 95% CI [0.72, 0.86]. Figure 4.3 depicts this correlation.

4.4.5.2 *Inferences from a single trial*

As detailed in the methods section for Experiment 4.1, we chose towers such that the evidence for the correct mass ratio was maximized. Consequently, all the towers we chose had the feature that if the tower fell under one mass ratio, it would *not* fall under the other, and vice versa. By considering this counterfactual, the observer gains more information than they otherwise would have; thus, Equation 4.1 does not tell the whole story. Let $\mathcal{L}(\kappa = k) := p(F_t|S_t, \kappa = k)$ be either

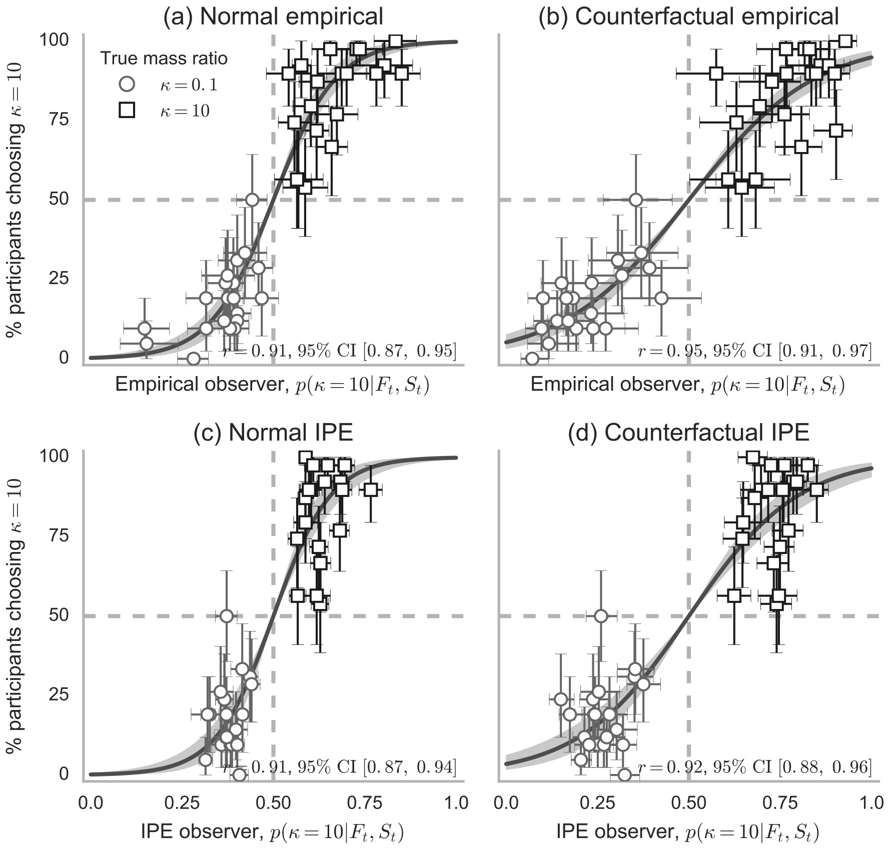


Figure 4.4: Comparing empirical and IPE likelihoods with and without counterfactual adjustments. Each subplot shows the posterior probability of $\kappa = 10$ in comparison to responses to ``which is the heavy color?'' in Experiment 4.1 according to different methods of calculating the likelihood. The best fitting sigmoid function for each relationship is plotted in black, with the shaded region indicating 95% confidence intervals for the best-fit coefficient. Error bars are bootstrapped 95% confidence intervals, and dotted lines show decision boundaries. (a) The posterior calculated using the original empirical likelihood. (b) The posterior calculated using the counterfactual empirical likelihood. (c) The posterior calculated using the original IPE likelihood. (d) The posterior calculated using the counterfactual IPE likelihood.

the IPE likelihood or the empirical likelihood. Then, the "counterfactual likelihood" is:

$$p_{\mathrm{CF}}(F_t|S_t, \kappa = k) = \frac{\mathcal{L}(\kappa = k)(1 - \mathcal{L}(\kappa \neq k))}{\mathcal{L}(\kappa = k)(1 - \mathcal{L}(\kappa \neq k)) + (1 - \mathcal{L}(\kappa = k))\mathcal{L}(\kappa \neq k)} \tag{4.9}$$

The derivation of this equation is given in Appendix A.4.

Figure 4.4a compares people's judgments with the posterior distribution calculated with the original empirical likelihood, and shows a clear sigmoidal relationship between the model and people, defined as $y = 1/(1 + \exp(-\beta(x - 0.5)))$. Using the original empirical likelihood, the best fit coefficient for sigmoid is $\beta = 11.75$, 95% CI $[9.82, 14.48]$. In contrast, when we switch to using the counterfactual likelihood (Figure 4.4b), this sigmoidal relationship significantly lessens ($\beta = 5.82$, 95% CI $[4.99, 6.84]$). To determine how different these sigmoid relationships are from linear, we computed best-fit sigmoid coefficients for linear data with Gaussian noise, with the variance of the noise equal to the variance in the residuals between people's judgments and the likelihood. We find that a truly linear relationship gives a best-fit coefficient of $\beta = 4.94$, 95% CI $[3.55, 6.81]$ when using the residual variance for the original likelihood, and $\beta = 4.92$, 95% CI $[4.05, 5.97]$ when using the residual variance for the counterfactual likelihood, indicating that the original likelihood is distinguishable from a linear relationship while the counterfactual likelihood is not. Thus, it appears that participants picked up on this counterfactual information and exploited it when making their inferences about the mass ratio. To account for this, all results reported from this point were computed using an IPE observer model and an empirical observer model that took this counterfactual information into account.

As depicted in Figure 4.4d, the correlation between the IPE observer model probabilities and average human judgments of the mass ratio was $r = 0.93$, 95% CI $[0.88, 0.96]$; for the empirical observer model (Figure 4.4b), it was $r = 0.95$, 95% CI $[0.91, 0.97]$. The correlation between the IPE observer model accuracy and human accuracy was $r = 0.25$, 95% CI $[-0.04, 0.51]$, and for the empirical observer model, it was $r = 0.60$, 95% CI $[0.36, 0.78]$. Note that the split-half correlation of accuracy (described in the results section of Experiment 4.1) was $r = 0.69$, 95% CI $[0.55, 0.80]$; thus, the empirical observer model was nearly at ceiling performance in predicting the level of agreement amongst participants. We emphasize here that because the empirical observer model is computed from participants responses to "will it fall?", our results show that people's judgments of relative mass can be predicted by their judgments of the towers' stabilities. This implies that there is a systematic relationship between the mechanism that people use both to make judgments of stability *and* judgments of relative mass.

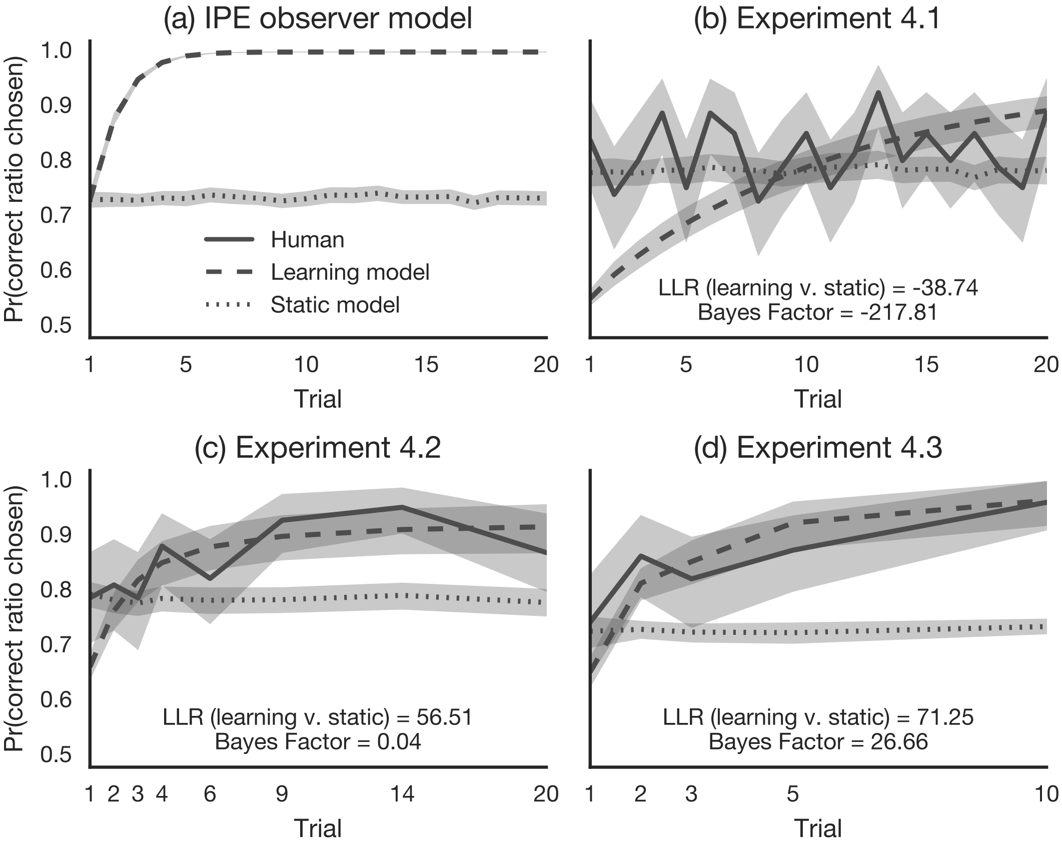


Figure 4.5: Model predictions of mass inferences. In all plots, dotted and dashed lines indicate the mean proportion of correct model responses computed for the IPE observer (responses for the empirical observer are nearly identical). Shaded regions are 95% confidence intervals of the mean. (a) The top left subplot shows the IPE learning and static observer models before being fit to human data (i.e., ideal observer predictions where $\beta = 1$ for all participants). (b) The static model is a better fit to the human data from Experiment 4.1 than the learning model. (c-d) The learning model is a better fit to human data from Experiments 4.2-4.3 than the static model.

4.4.5.3 *Learning over multiple trials*

Figure 4.5 shows a comparison of the human data to the fitted models under the IPE observer, and Table 4.2 lists numerical values for the log-likelihood ratios and Bayes factors. Negative values indicate evidence for the static model, and positive values indicate evidence for the learning model. In Experiment 4.1, the static model was a better explanation for people's behavior than

Table 4.2: Log-likelihood ratios (LLR) and Bayes factors ($\log K$) for learning vs. static models. Positive values favor the learning model, while negative values favor the static model.

		Exp. 4.1	Exp. 4.2	Exp. 4.3 within subjs.	Exp. 4.3 between subjs.
LLR	IPE (fitted)	-38.74	56.51	43.42	71.25
	Empirical (fitted)	-70.79	60.02	41.77	60.78
$\log K$	IPE	-217.81	0.04	-0.52	26.66
	Empirical	-269.85	-6.58	-9.78	22.15

the learning model. In Experiment 4.2, the learning model was the better explanation of participant's behavior than the static model according to the log-likelihood, though not according to the Bayes factor. Similarly, if we look only at the condition in which participants responded on five trials in Experiment 4.3 (within subjects), the learning model was a better explanation of people's behavior according to the log-likelihood ratio, but not the Bayes factor. If we look between subjects in Experiment 4.3, using data from only the first response of each subject, we find that the learning model is a better explanation of people's behavior according to both measures. This result is consistent with the Spearman rank correlations reported in the results section of Experiment 4.3.

Why do the log-likelihood ratios and Bayes factors disagree on the results for Experiment 4.2 and within-subjects in Experiment 4.3? This disagreement illuminates the surprising results we obtained in Experiment 4.2: while the majority of participants did seem to take into account the evidence and learn over time, there was a minority of participants who seemed to be answering randomly. Figure 4.6 shows the distribution of fitted parameters for both the static and learning models (computed with the IPE likelihood) in all three experiments. The distributions resulting from using the empirical likelihood are very similar. Unsurprisingly, the parameters for the learning model in Experiment 4.1 are clustered around $\beta = 0$, indicating that the evidence is largely ignored. The parameters for the static model in Experiments 4.2-4.3 are biased towards $\beta > 1$, indicating that the evidence is weighed more strongly (which is what we would expect if people were getting more accurate over time). The parameters for the learning model in Experiments 4.2-4.3 indicate a small fraction of participants who seemed to mostly ignore the evidence. In contrast, the majority of participants in Experiments 4.2-4.3 have $\beta = 1$, indicating that they are best fit to the ideal observer model.

Based on the histograms in Figure 4.6, we can understand the disagreement between the log-

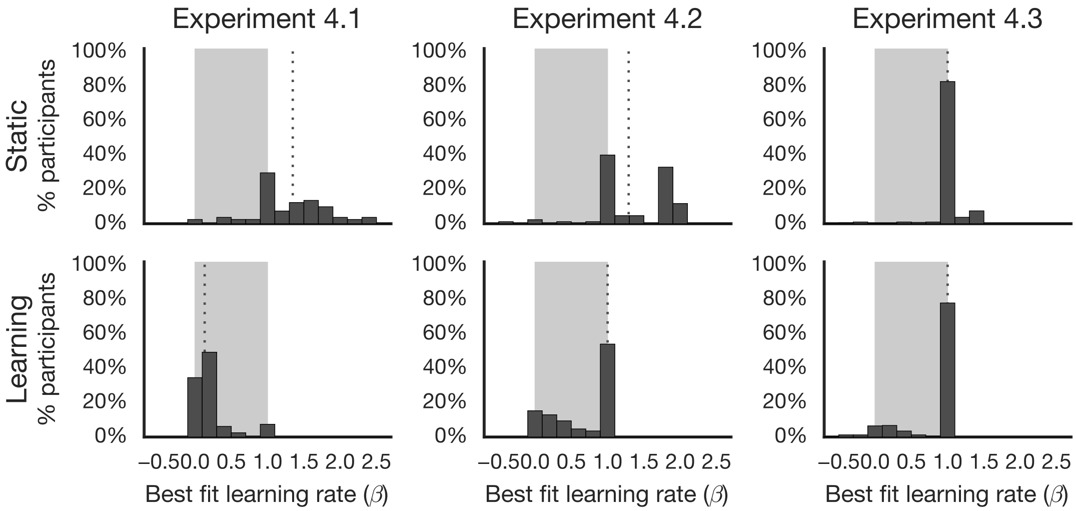


Figure 4.6: Distributions of parameter estimates. Each model was fit separately to each participant in each experiment. The top row shows the distributions for parameters fitted to the static model, while the bottom row shows the distributions for the learning model. The shaded gray area in each figure shows $0 < \beta < 1$, and the dotted lines show the medians of each histogram. If $\beta = 1$, the fitted model was exactly the same as the ideal observer model. If $\beta > 1$, the fitted model weighed evidence more strongly than the ideal observer. If $0 < \beta < 1$, the fitted model weighed evidence less strongly than the ideal observer. If $\beta = 0$, the fitted model did not take evidence into account at all. Finally, if $\beta < 0$, the fitted model took into account the *opposite* of what the evidence suggested.

likelihood ratios and Bayes factors in Experiment 4.2 and within-subjects in Experiment 4.3. When β is fit to each participant individually in the learning model, the resulting fits are close to zero for the participants who seemed to be ignoring the evidence. The behavior of these participants is close to random, and is therefore equally well explained by both models when $\beta \approx 0$. However, the majority of participants did actually learn over time, and had fitted coefficients of $\beta = 1$ for the learning model. The static model does a poor job of explaining the responses of these participants. Thus, overall the learning model is favored when we look at log-likelihood ratios resulting from fitted parameters.

The Bayes factors disagree with the log-likelihood ratios, however, because the fits under the learning model for people who did not learn ($\beta \approx 0$) have relatively low prior probability under the Laplace prior (with parameters $\mu = 1$ and $b = 1$). So, even though these coefficients result in higher log-likelihood ratios, they do not get much weight when the Bayes factors are computed.

For the non-learning participants, the coefficients that do have higher prior probability yield responses that have low probability under the learning model and higher probability under the static model. This is enough to outweigh the fact that the learning model is a good model for people who were in fact learning, thus resulting in Bayes factors that favor the static model in Experiment 4.2 and within-subjects in Experiment 4.3.

4.5 General Discussion

We asked whether people can infer unobservable physical properties in complex scenes. We ran three experiments to answer this question, and proposed a new class of models that can capture this phenomenon by combining probabilities generated from simulations with Bayesian inference. In Experiment 4.1, we found that across participants and stimuli, 80% of the judgments of relative mass were correct, even though these inferences were informed by only a single trial's worth of information. This result contradicts previous studies suggesting that people are poor at inferring hidden properties like mass (e.g. Gilden & Proffitt, 1989a, 1989b) in all but the simplest of scenarios. If anything, it suggests that people are remarkably *good* at inferring such properties. Moreover, Experiments 4.2-4.3 suggest that people can accumulate information and become increasingly fine-tuned to the properties of their environment the longer they observe it.

As illustrated by Figure 4.4b, we also found a systematic relationship between people's predictions about the future of a physical scene and their inferences about that scene's parameters. This suggests that people make predictions about whether the tower will fall under different possible parameter values in order to infer which parameter values were more likely to have produced the actual outcome they observed. This supports the idea that people rely on the same mechanism or source of physical knowledge both for predicting the future of physical scenes, and for inferring properties about objects in those scenes.

In the remainder of this article, we discuss questions for future research, as well as how these results relate to the larger literatures on physical reasoning.

4.5.1 Can people make more detailed inferences?

One question for future research regards whether people can make more detailed inferences beyond the ratios that were used in the present experiments. While within the range of objects that people normally interact with, the ratios of 10:1 and 1:10 that we used are relatively high. Can people distinguish between smaller mass ratios as well? Moreover, can people infer the *specific* ratio from a number of alternatives (e.g., 1:2 vs. 1:3 vs. 1:4)? We speculate that people could probably tell the difference between very small and very large mass ratios (e.g., 1:2 vs. 1:10) but have difficulty distinguishing between similar mass ratios (e.g., 1:2 vs. 1:3). In terms of whether people can infer a specific ratio, we suspect that in general people will be good at determining whether one type of object is heavier than other, but may have difficulty determining precisely by how much.

4.5.2 How do size and material properties influence people's inferences of mass?

In the present experiment, we asked participants to reason about the relative mass of objects. However, our stimuli consisted of blocks that were all the same size, and which were all subject to the same gravitational acceleration; thus, it could be that what we were measuring was not people's ability to reason about mass but rather an ability to reason about density or weight. An important direction for future research will be to disentangle these confounding factors.

If people reason are sensitive to both mass and density, then their responses should influenced by factors such as size and material properties. There is ample evidence that people have strong expectations about physical properties based on their size and perceived material. Expectations about material densities are generally correct and lead to accurate predictions about the stability of objects (Lupo & Barnett-Cowan, 2015). Thus, we would expect people to be biased in their inferences if visual cues provide evidence for different sizes or materials.

It is not *a priori* clear in what direction people's inferences should be biased, however. For example, the size-weight illusion (SWI) is a well documented phenomena in which people perceive the smaller of two equally-weighted objects to be heavier after lifting them (Charpentier, 1891; Flanagan & Beltzner, 2000; Flanagan, Bittner, & Johansson, 2008; Grandy & Westwood, 2006; Murray, Ellis, Bandomir, & Ross, 1999; H. E. Ross, 1969). The explanation for this effect is that people expect the smaller object to be lighter than it actually is, and consequently perceive it to be heavier. Would visual evidence—rather than sensorimotor evidence—produce the same effect, in which people perceive smaller objects to be heavier when they are actually the same weight, because the visual evidence is surprising?[5] Or, would people's prior expectations simply override the evidence, causing them to perceive the smaller objects as lighter?

In a similar vein, the material-weight illusion (MWI) is another effect in which people per-

[5] For example, if the observer expects the smaller object to be lighter, then the observer would expect the smaller object to move faster than the larger object after a collision in which the initial velocities are the same. However, if they are actually the same weight, then they will move at the same speed after the collision.

ceive objects of heavier-looking materials (e.g., metal) to be lighter than objects of lighter-looking materials (e.g., styrofoam) even when those objects are actually the same weight (Buckingham, Ranger, & Goodale, 2011; Ellis & Lederman, 1999; Seashore, 1899; Wolfe, 1898). We can ask the same question: would this effect persist when people only have visual evidence, or would expectations override the visual evidence? Based on our experience with blocks of different materials (see Battaglia et al., 2013), we hypothesize that expectations would tend to overrule the visual evidence (i.e., that people would report objects with heavier-looking materials to be heavier, even when they are not).

4.5.3 Can these results be reconciled with "naïve physics" errors?

Many studies suggest that people's ability to reason about the physical world is relatively impoverished and error-prone. For example, some researchers have argued that people can reason about only a single dimension of physical information at once (e.g. the position of the object's center of mass over time) and have trouble incorporating multiple dimensions of information (e.g. mass as well as position) (Gilden & Proffitt, 1989b). Rather than accurately incorporating these multiple dimensions of information, Gilden and Proffitt (1989a) suggested that when reasoning about mass, people rely only on simple heuristics such as that the faster object after a collision is the lighter object. Todd and Warren (1982) also suggested that people rely on limited information (such as the final speeds of the objects) which is accurate in certain cases, but not in all situations (such as when elasticity is very high or low). While varying accounts of different heuristics and biases seem to conflict with each other in which effects they can predict, they can be reconciled by the noisy Newton hypothesis. Specifically, these effects can be explained as being the result of perceptual uncertainty (Sanborn, 2014; Sanborn et al., 2013). Moreover, given the results of the present paper, it seems unlikely that people only pay attention to a single dimension of information when making inferences about mass: in our experiments, participants must pay attention to the position, orientation, and color of multiple three-dimensional objects.

Other research has shown that people sometimes seem to rely on a naïve theory of "impetus", resulting in incorrect beliefs such as that if someone drops an object while they are walking, the object will fall straight down (rather than in a parabolic curve due to the combination of horizontal and vertical velocity) (McCloskey, 1983; McCloskey et al., 1983). McCloskey et al. (1983) argued that this particular effect is due to a perceptual illusion involving the frame of reference of the motion (that is, that the object appears to move straight down with respect to the motion of the carrier). This interpretation is not inconsistent with the approximate simulation hypothesis, however: if approximate simulations are learned from perception, then perceptual illusions *should* affect the resulting dynamics models.

Other errors documented in the naïve physics literature may be a result of engaging different forms of physical knowledge. One classic error involves a pendulum task, in which a participants are asked to consider a pendulum consisting of a bob on a string. They are told that the string breaks, and are asked to draw the resulting trajectory of the bob. Participants' responses to this task tend to be inconsistent and often incorrect. However, Smith, Battaglia, and Vul (2013) gave people analogous tasks of *catching* the bob in a bucket or *cutting* the string such that the bob would go into a fixed bucket, and found that people were quite accurate in these tasks. When their responses did deviate from ground truth, they were strongly predicted by a noisy Newton simulation model.

Perhaps, then, certain tasks engage accurate predictive knowledge of object dynamics, while others engage more error-prone conceptual knowledge. This hypothesis has previously been suggested by Schwartz and Black (1999), who found a similar dissociation between explicit conceptual knowledge and implicit motor knowledge. Schwartz and Black (1999) asked participants to judge which of two glasses of water would need to tilt further before the water reached the rim of the cup. When given this judgment explicitly, participants were overwhelmingly incorrect; however, when asked to tilt an empty cup and *imagine* the water reaching the rim of the glass, participants tilted the cups to the appropriate angle. P. A. White (2012) further discusses the dissociation between the people's accuracy in tasks that seem to engage motor knowledge versus their inaccuracy in tasks that seem to only engage visual knowledge. Similarly, representational momentum tasks seem to engage a type of physical knowledge that is dissociated from explicit formal knowledge of physics (Freyd & Jones, 1994; Kozhevnikov & Hegarty, 2001).

A pertinent question is: why do some tasks seem to invoke one system of knowledge, while others invoke another? It is not the case that people only use motor knowledge when the task calls for taking actions, and only visual knowledge when the task does not. For example, the classic mental rotation task by Shepard and Metzler (1971) is a good example of what seems like a purely perceptual task, yet that still engages the motor system (Parsons, 1994). One potential answer to this question is that the mind attempts to find a "perceptual match" to stored representations of actions on objects, and falls back on purely visual knowledge only when there is no motor representation to be found (P. A. White, 2012). An alternate hypothesis is that the mind engages in a metacognitive task of strategy selection (e.g. Lieder et al., 2014) and picks the strategy or domain of knowledge that has the higher expectation of being useful in the given situation.

4.5.4 How does approximate physical simulation relate to internal models for sensorimotor and perceptual prediction?

It is critical for the motor system to be able to accurately predict and respond to novel object dynamics; if it could not do so, then we would be unable to effectively interact with those objects. Indeed, there is a wealth of literature on motor and perceptual learning that suggests that people are quite sensitive to many dimensions of physical information and that they take this information into account in a reasonable way. There is evidence that the sensorimotor system learns both inverse models of control (i.e., what actions to perform to achieve a particular state) as well as forward models of interactions with the environment (i.e., what state to expect when an action is taken in the current state) (Flanagan, Vetter, Johansson, & Wolpert, 2003; Kawato, 1999; Wolpert & Kawato, 1998). These models incorporate dynamical information along multiple dimensions; for example, by including both gravity and mass, the motor system can predict the momentum of an oncoming object and appropriately contract the relevant muscles in order to catch it (Zago & Lacquaniti, 2005).

The types of approximate physical simulations hypothesized in this paper and by Battaglia et al. (2013) are a type of forward dynamics model: given the current state, they predict the next state. Of course, it is not clear whether the forward models used by the motor system are exactly the same as those engaged in the higher-level cognitive tasks investigated here, especially given that our tasks did not involve a motor component. Given the evidence for multiple forward and inverse models in the motor system (Wolpert & Kawato, 1998), it seems likely that there could be additional forward models used by other aspects of cognition such as the perceptual system. Moreover, there is good evidence to believe that there is at least some degree of decoupling between the motor and perceptual systems. For example, illusions such as the size-weight illusion and material-weight illusion tend to persist, even after the motor system has adapted (Flanagan & Beltzner, 2000; Grandy & Westwood, 2006), and are influenced by top-down knowledge (Ellis & Lederman, 1998).

If the motor system is not directly involved in the approximate physical simulations explored in this paper, what is? An alternate explanation might be the perceptual forward models underlying an interesting class of phenomena known as *displacement* or *representational momentum* effects (Freyd & Finke, 1984; Hubbard, 2005). In these experiments, people's memory for the location of an object is distorted in the direction of implied motion: if someone sees an object moving towards the left, then they remember the object as being slightly more to the left than it actually was. These effects extend beyond just objects *in* motion but also to static objects that *could* move, such as an object that would fall due to gravity or be pushed upwards by a spring (Freyd et al., 1988). In fact, displacement effects have been found involving many types of physical information, including friction, linear velocity, centripetal force, barriers, and even top-down knowledge of properties such as mass (see Hubbard (2005) for a review).

These results from the literature on displacement suggest that the perceptual system has a rich and detailed knowledge of how objects behave. Yet, it also appears that displacement effects are more consistent with the impetus theory of physical reasoning than with Newtonian physics. For example, there are greater displacements for objects moving up when the objects are small rather than when they are large (Kozhevnikov & Hegarty, 2001); this is similar to the Aristotelean belief that heavier objects fall faster than lighter objects. Similarly, for objects moving in a curved path, there are displacements in the direction of centripetal force (Freyd & Jones, 1994; Hubbard, 1996); this result is similar to the naïve belief that objects moving in a curved tube will continue moving in a curved path after exiting the tube (McCloskey & Kohl, 1983). Other evidence similarly suggests the presence of impetus principles underlying displacement effects in Michotte-type launching experiments (Hubbard, 2004, 2013a, 2013b, 2013c; Hubbard & Ruppel, 2002).

There are two hypotheses that come to mind in reconciling approximate physical simulation with displacement. The first hypothesis is that approximate physical simulation is a different cognitive process from that underlying displacement. This argument is plausible, though lacks in parsimony as it requires positing that the mind has two forward models for the same physical phenomena. The second hypothesis is that approximate physical simulation and displacement are related, and that the impetus effects arise naturally from learning forward dynamics from noisy data. Whether either of these hypotheses is correct is an open question for future research.

4.5.5 Is simulation too computationally intensive?

The noisy Newton hypothesis is largely posited at the computational-level of analysis (Marr, 1982), while approximate physical simulation is an algorithmic-level solution to the computational-level problem. As an algorithmic-level model, approximate physical simulation comes with several dimensions of computational constraints that allow us to form testable hypotheses about how it is used. One question we can ask is: given the ability to run approximate physical simulations, how many simulations should be run? In particular, one critique of approximate physical simulation is that it is too computationally intensive to be plausible as a cognitive mechanism (Davis & Marcus, 2014). It seems unlikely that people run hundreds or even tens of simulations per decision; a more realistic hypothesis is that people run just one or two simula-

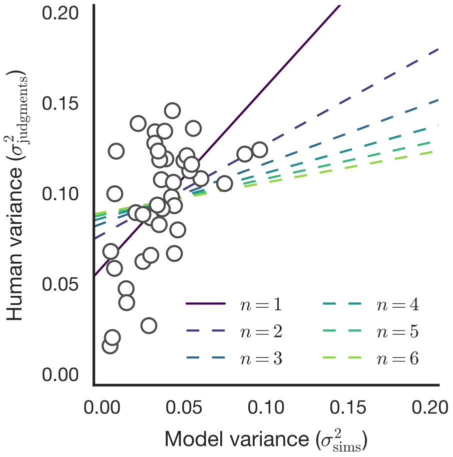


Figure 4.7: Analysis of human variance. This plot shows the variance of human predictions as a function of the variance of model predictions computed from different numbers of samples (n). The solid line indicates the fit with the lowest mean squared error, corresponding to $n = 1$.

tions per decision.

Battaglia et al. (2013) performed an analysis of the standard deviation of participants' "will it fall?" judgments to estimate the number of simulations. We perform a similar analysis here. If people take n samples per judgment, then the variance of their responses should be:

$$\sigma^2_{\mathrm{judgments}} = \frac{\sigma^2_{\mathrm{sims}}}{n} + \omega^2, \tag{4.10}$$

where σ_{sims} is the standard deviation of simulations from the IPE and ω is the standard deviation of other sources of uncertainty (such as general decision-making noise). We used a least-squares linear regression to estimate ω in Equation 4.10 for each of $n \in \{1, 2, 3, 4, 5, 6\}$, and from each regression computed the mean squared error between predicted variance and the actual variance of human judgments. Figure 4.7 shows the variance of participant responses as a function of the variance of IPE samples. The dotted lines correspond to the predicted variance for each value of n. The solid line has the lowest MSE, and corresponds to $n = 1$ and $\omega = 0.25$, indicating that the variance of participants' judgments is consistent with them running one simulation per judgment. This result is consistent with research suggesting that it is optimal to only take a small number of samples before making a decision (Vul et al., 2014), and that this is in fact what people do in the case of physical prediction (see Chapter 6).

Running a single simulation per judgment seems much more tractable than, say, ten simulations. However, there are additional questions regarding computational limitations beyond just the number of simulations: for example, it seems unlikely that people would be able to run a detailed simulation of ten objects, especially given that people can only track a small number of objects simultaneously (Pylyshyn & Storm, 1988). While it is beyond the scope of this paper to answer this question, we emphasize that our hypothesized simulations are *approximate* and thus may sacrifice accuracy for performance.[6] How this trade-off actually manifests is an important question that we hope to address in the future.

While we have argued that there are situations in which simulation is computationally tractable—such as the prediction and inference tasks presented in this paper—there are certainly cases where simulation is always going to be too expensive. For example, Smith, Dechter, et al. (2013) describe a task in which participants have to predict whether a ball will reach a green target or a red target first. Interestingly, while some of their stimuli seem to lend themselves to a physical simulation, others clearly do not (such as if the ball is in a box with the green target in the box and with the red target outside of the box). In such cases, it may be that conceptual or qualitative knowledge is more appropriate (Forbus, 1983, 2011). How the mind decides between these approaches could perhaps be thought of as another instance of metacognitive strategy selection (e.g. Lieder et al., 2014).

4.5.6 Conclusion

In sum, we found that people can learn the relative masses of objects after observing their interactions in complex scenes. Our results both confirm recent reports that people's physical scene understanding is driven by probabilistic physical simulation, and provide additional evidence that the *same* mechanism used in making predictions is also used when making inferences about underlying physical parameters. Beyond simply making one-shot inferences, our results suggest how mental simulation as a cognitive resource can also be used to aggregate evidence and learn about properties of the world over time. Mental simulation thus truly offers an "infinite use of finite means" by simultaneously serving the needs of prediction, inference, and learning. Yet, there remain a number of unanswered questions: how do people know what simulations to run in the

[6] We distinguish our approximate simulations from heuristics in that heuristics are typically thought of as simple rules that tend to do a good job, but which are not necessarily approximating an optimal solution. Although some research suggests that some heuristics may indeed be approximating an optimal solution (Lieder, Griffiths, & Goodman, 2012; Lieder et al., 2014; Tenenbaum & Griffiths, 2001), this has not been demonstrated for *all* heuristics, and thus we feel it is more informative to label our approach as an approximation to an optimal solution rather than as a heuristic.

first place, and how many simulations should be run? In the next chapters, we will investigate the answers to these questions in further detail.

Think left and think right and think low and think high. Oh, the thinks you can think up if only you try!

Dr. Seuss, *Oh, the Thinks You Can Think!*

5
Selecting Computations

One of the most astonishing cognitive feats is our ability to envision, manipulate, and plan with objects—all without actually perceiving them. The previous chapter illustrated a striking example of this, demonstrating that people can use mental simulation both to make predictions of the physical world (e.g., whether a tower of blocks will fall) as well as inferences of unobservable properties (e.g., which blocks are heavier than others). However, such demonstrations of mental simulation fail to address one of the most fundamental questions about it: how people decide *what* to simulate.

Mental rotation provides a simple example of the decision problem posed by simulation. In the classic experiment by Shepard and Metzler (1971), participants viewed images of three-dimensional objects and had to determine whether the images depicted the same object (which differed by a rotation) or two separate objects (which differed by a reflection and a rotation). They found that people's response times (RTs) had a strong linear correlation with the minimum angle of rotation, a result which led to the conclusion that people solve this task by "mentally rotating" the objects until they are congruent. However, this explanation leaves several questions unanswered. How do people know the axis around which to rotate the objects? If the axis is known, how do people know which direction to rotate the objects? And finally, how do people know how long to rotate?

The text of this chapter was previously published as Hamrick and Griffiths (2014).

Previous models of mental rotation have largely focused on the representation of mental images, rather than how people decide *which* mental images to generate. Kosslyn and Shwartz (1977) proposed a model of the mental imagery buffer, but did not say *how* it should be used. Similarly, Julstrom and Baron (1985) and Glasgow and Papadias (1992) were mostly concerned with modeling the representational format underlying imagery. Although Anderson (1978) emphasized the importance of considering both representation and process, he dismissed the problem of determining the direction of rotation as a "technical difficulty".

The only models that have seriously attempted to address the decision of *what* to simulate are those by Funt (1983) and Just and Carpenter (1985). In both of these models, the axis and direction of rotation are computed prior to performing the rotation. One object is then rotated through the target rotation, and is checked against the other object for congruency. However, this approach assumes that the corresponding points on the two objects can be easily identified, which is not necessarily the case. Additionally, recent research shows that when performing *physical* rotations, people do not rotate until congruency is reached; they may even rotate *away* from near perfect matches (Gardony et al., 2014).

If people are not computing the rotation beforehand, what might they be doing? To answer this question, we perform a rational analysis of the problem of mental rotation (Anderson, 1990; Marr, 1982; Shepard, 1987). At the computational level, we can say that the *problem* is to determine which spatial transformations an object has undergone based on two images of that object (which do not include information about point correspondences). At the algorithmic level, we are constrained by the notion that mental images must be transformed in an analog manner (or in a way that is approximately analog), and that mental images are time-consuming and effortful to generate. Thus, the *goal* is to make this determination while performing a minimum amount of computation (i.e., as few rotations as possible).

The original "congruency" hypothesis (Shepard & Metzler, 1971) is a rational solution to this problem, in the sense that the smallest amount of computation coincides with rotating through the minimum angle. However, it violates the constraint that we do not know the points of correspondence between the images, which is what necessitates the use of imagery. Noting that a rational solution need not maintain a single trajectory of rotation, we explore an alternative model, which—rather than computing the angle of rotation—engages in an *active sampling* strategy.

Active sampling is the idea that people gather new information in a manner that increases certainty about the problem space. An everyday example of this can be observed in the game of "20 questions", in which one person thinks of a concept, and another has to guess the concept in 20 questions or less. The first question is almost always "person, place, or thing?", because the answer provides the most possible information about the concept of interest. Active sampling has gained support across several areas of cognitive science (e.g. Gureckis & Markant, 2012), including other spatial domains (Juni, Gureckis, & Maloney, 2011). In the case of mental rotation, actively choosing rotations may be the best way to gather evidence about the similarity between the observed shapes when the angle of rotation is unknown.

In this paper, we explore these questions through rational analysis (Anderson, 1990; Marr, 1982; Shepard, 1987) and compare four models of mental rotation. We begin the paper by discussing the previous literature on mental imagery. Next, we outline computational- and algorithmic-level analyses of the problem of mental rotation. We then describe a behavioral experiment based on the classic mental rotation studies (e.g. Cooper, 1975), and compare the results of our experiment with each of the models. We conclude with a discussion of the strengths and weaknesses of each model, and lay out directions for future work.

5.1 Modeling Mental Rotation

In this section, we formalize our rational analysis and propose four models of mental rotation: one based on existing models; two which are extensions of the first but with relaxed assumptions; and one based on the active sampling approach.

The task we are interested in modeling involves observing two images and determining whether one image depicts the "same" object as the other image (differing by a rotation), or a "flipped" version of the object in the other image (differing by a reflection and then a rotation).

5.1.1 Computational-level analysis

We denote the shapes as X_a and X_b and assume X_b is generated by a transformation of X_a, i.e. $X_b = f(X_a, \theta, h)$, where θ is a rotation, $h = 0$ is the hypothesis that the images depict the same object, and $h = 1$ is the hypothesis that the images depict mirror-image objects. The posterior probability of each hypothesis given the observed shapes is then:

$$p(h \mid X_a, X_b) \propto \int p(X_b \mid X_a, \theta, h)p(h)p(\theta)\,\mathrm{d}\theta, \tag{5.1}$$

where $p(X_b \mid X_a, \theta, h)$ is the probability that X_b was generated from X_a. Because we want to determine which hypothesis is more likely, the quantity of interest is a posterior odds ratio $\mathcal{B} := p(h = 0 \mid X_a, X_b)/p(h = 1 \mid X_a, X_b)$ which (assuming that all rotations are equally likely) is equivalent to:

$$\mathcal{B} = \frac{\left(\int p(X_b \mid X_a, \theta, h = 0)\,\mathrm{d}\theta\right) \cdot p_0}{\left(\int p(X_b \mid X_a, \theta, h = 1)\,\mathrm{d}\theta\right) \cdot p_1}, \tag{5.2}$$

where $p_0 = p(h = 0)$ and $p_1 = p(h = 1)$, for brevity. If $\mathcal{B} > 1$, then we accept the hypothesis that the images depict the same object ($h = 0$); if $\mathcal{B} < 1$, then we accept the hypothesis that the images depict flipped objects ($h = 1$).

5.1.2 Algorithmic constraints

We represent a shape of N vertices with a $N \times 2$ coordinate matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, and denote the rotation anor reflection transformation as $f(\mathbf{X}, h, \theta) := \mathbf{X}\mathbf{F}_h^T\mathbf{R}_\theta^T$, where $\mathbf{R}_\theta$ is a rotation matrix, and $\mathbf{F}_h$ is either the identity matrix $\mathbb{I}$ (when $h = 0$) or a reflection matrix across the y-axis (when $h = 1$).

We define $p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h)$ to be the similarity between $\mathbf{X}_b$ and a transformation of $\mathbf{X}_a$: $p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h) := S(\mathbf{X}_b, f(\mathbf{X}_a, h, \theta))$. We do not know which vertices of $\mathbf{X}_b$ correspond to which vertices of $\mathbf{X}_a$, so the similarity S must marginalize over the set of possible mappings. For brevity, let $\mathbf{X}_m = \mathbf{M} \cdot f(\mathbf{X}_a, h, \theta)$ where $\mathbf{M}$ is a permutation matrix. Then:

$$S(\mathbf{X}_b, f(\mathbf{X}_a, h, \theta)) := \frac{1}{2N} \sum_{\mathbf{M}} \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_{bn} \mid \mathbf{x}_{mn}, \mathbb{I}\sigma_S^2), \tag{5.3}$$

where $2N$ is the total number of possible mappings,[1] and $\sigma_S^2 = 0.15$ is the variance of the similarity. Example similarity curves are shown in Figure 5.1.

We assume that the observed shapes must be transformed by a small amount at a time, and each transformation takes a non-negligible amount of time. If the current mental image is $\mathbf{X}_t$, then:

$$\mathbf{X}_{t+1} = \begin{cases} f(\mathbf{X}_t, 0, \epsilon) & \text{rotate by } \epsilon \text{ radians,} \\ f(\mathbf{X}_t, 1, 0) & \text{flip,} \\ f(\mathbf{X}_a, 0, 0) & \text{reset to } 0^\circ\text{, or} \\ f(\mathbf{X}_a, 1, 0) & \text{reset and flip,} \end{cases} \tag{5.4}$$

where $\epsilon \sim \left|\mathcal{N}(0, \sigma_\epsilon^2)\right|$ and σ_ϵ^2 is the variance of the step size.

[1] It is $2N$ and not N^2 because, in polar coordinates, vertices are always connected to their two nearest neighbors in the θ dimension.

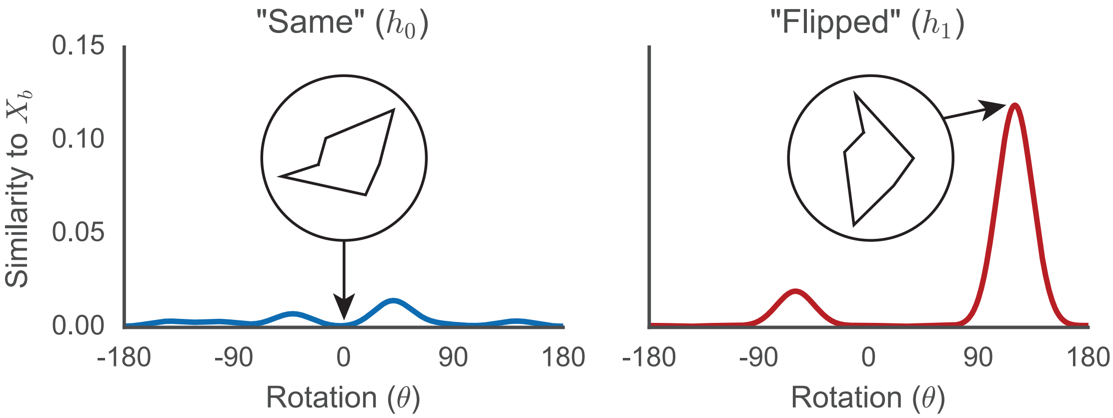


Figure 5.1: Example stimuli and similarities. This figure shows a ``flipped'' stimulus pair with a rotation of $120°$, and the corresponding similarity functions for each hypothesis. Arrows indicate where each shape lies on the curve.

To summarize, we approximate the likelihood term of Equation 5.2 using the similarity function defined in Equation 5.3. Because we assume mental rotations are performed sequentially, this similarity can only be computed for the actions listed in Equation 5.4.

5.1.3 Specific models of mental rotation

In order to approximate Equation 5.2 using samples of the similarity function, we must decide *which* places to sample and *when* stop sampling. The models below differ in how they make these decisions.

5.1.3.1 Oracle model

One hypothesis is that people compute the direction and extent of rotation beforehand using *a priori* knowledge of the correspondence between points in the images (Funt, 1983; Just & Carpenter, 1985). To reflect this hypothesis, we created an "oracle" model which is told which points on each shape correspond. From that correspondence, it computes the correct rotation and rotates through it.

To determine the correct rotation, we solve for the rotation matrix by computing $(\mathbf{X}_a\mathbf{F}_h^T)^{-1}_{\text{left}} \cdot \mathbf{X}_b$, where $(\mathbf{X}_a\mathbf{F}_h^T)^{-1}_{\text{left}}$ is the left inverse of $\mathbf{X}_a\mathbf{F}_h^T$. We then check each h to see if the computation produces a valid rotation matrix; the h that does is the correct hypothesis. This gives us the true value of θ, so Equation 5.2 becomes a generalized likelihood ratio test, where θ is set to the MLE value, rather than being marginalized:

$$\mathcal{B} = \frac{\max_\theta p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h=0) \cdot p_0}{\max_\theta p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h=1) \cdot p_1}. \tag{5.5}$$

If we give equal weight to the two hypotheses, then the priors cancel out; if we weigh one hypothesis more heavily, then our decision will be biased towards that hypothesis. However, unless the likelihood ratio is already very close to 1, small biases in the prior will not make much of a difference.

5.1.3.2 *Threshold model*

A model which does not know point correspondences could use the following algorithm: (1) pick a random direction; (2) take a single step; (3) if that step decreased similarity, then begin rotating in the reverse direction, otherwise continue rotating in the original direction; (4) continue rotating in the chosen direction until a "match" is found (defined as finding a value of S that exceeds a threshold); and (5) if no match was found, flip, and start over from step one. We only allow for the "flip" action after no match has been found, because there is no particularly principled way for the Threshold model to choose when to flip. We assume that the locations where S is greater than the threshold correspond to the true θ (or points near the true θ). So, as with the Oracle model, we use Equation 5.5.

5.1.3.3 *Hill Climbing model*

In the current formulation of the problem, choosing the threshold is straightforward because we know both the exact geometry of the shapes and that a linear transformation exists which will align them. However, this choice is not always clear *a priori*, as the global optimum depends on many factors (e.g., shape complexity, dimensionality, perceptual uncertainty, and whether the shapes are identical). One way to deal with the problem of choosing a threshold would be use a global optimization strategy; however, this would not result in the linear RT found by Shepard and Metzler (1971). A second alternative is to perform a Hill Climbing (HC) search; i.e., rotate in the direction that increases similarity until no further improvement can be found. In contrast with the Threshold model, this results in arriving in a *local* maximum (which may or may not be the global maximum). Thus, as with the Oracle and Threshold models, we use Equation 5.5. We only allow for the "flip" action after a local maximum has been reached, because like the Threshold model, there is otherwise no principled way for the HC model to choose when to flip.

5.1.3.4 *Bayesian Quadrature model*

While the previous few models all focused on *searching* for the global maximum, we need only *approximate* Equation 5.2. We hypothesize a model based on the idea of *active sampling* (e.g. Gureckis & Markant, 2012): instead of searching for a maximum, we maintain a probability distribution over our *estimate* of Equation 5.2, and then sample actions which are expected to improve that estimate. This strategy has the benefits that it does not make assumptions about the scale of the similarity function; and, by choosing to sample places which are informative, this method implicitly minimizes the amount of rotation.

We denote Z_h as our estimate of the likelihood for hypothesis h, and write its expectation as:

$$\mathbb{E}[Z_h] = \int \left[\int S(\mathbf{X}_b, f(\mathbf{X}_a, \theta, h))p(\theta)\,\mathrm{d}\theta\right] p(S)\,\mathrm{d}S,$$

where S is the similarity function, and $p(S)$ is a prior over similarity functions. This method of estimating an integral is known in the machine-learning literature as *Bayesian Quadrature* (Diaconis, 1988; Osborne et al., 2012), or BQ. Denoting $S_h = S(\mathbf{X}_b, f(\mathbf{X}_a, \theta, h))$, we first place a *Gaussian Process* (Rasmussen & Williams, 2006), or GP, prior on the log of S_h in order to enforce positivity after it is exponentiated, i.e. $\mathbb{E}[Z_h] \approx \int \exp(\mu_h(\theta))p(\theta)\,\mathrm{d}\theta$, where $\mu_h := \mu(\log S_h)$ is the mean of the log-GP (Osborne et al., 2012). To approximate this integral, we fit a second GP over points sampled from the log-GP, which we denote as $\bar{S}_h := \exp(\mu_h)$. Then, from Duvenaud (2013), we have:

$$\begin{aligned}\mathbb{E}[Z_h] &\approx \int \bar{\mu}_h(\theta)p(\theta)\,\mathrm{d}\theta,\\ \mathbb{V}(Z_h) &\approx \iint \bar{\mu}_h(\theta)\mathrm{Cov}_h(\theta, \theta')\bar{\mu}_h(\theta')p(\theta)p(\theta')\,\mathrm{d}\theta\,\mathrm{d}\theta',\end{aligned}$$

where $\bar{\mu}_h := \mu(\bar{S}_h)$ is the mean of the second GP, and $\mathrm{Cov}_h := \mathrm{Cov}(\log S_h)$ is the covariance of the log-GP.

We approximate $p(Z_h) \approx \mathcal{N}(Z_h \mid \mathbb{E}[Z_h], \mathbb{V}(Z_h))$, which gives us a distribution over the likelihood ratio in Equation 5.2:

$$p(\mathcal{B}) \approx \frac{\mathcal{N}(Z_0 \mid \mathbb{E}[Z_0], \mathbb{V}(Z_0)) \cdot p_0}{\mathcal{N}(Z_1 \mid \mathbb{E}[Z_1], \mathbb{V}(Z_1)) \cdot p_1}.$$

This distribution cannot easily be calculated, but we are only interested in whether $Z_0 > Z_1$ or

$Z_1 > Z_0$. So, we use $Z_D = p_0 \cdot Z_0 - p_1 \cdot Z_1$ and compute

$$p(Z_D) \propto \mathcal{N}(p_0 \cdot \mathbb{E}[Z_0] - p_1 \cdot \mathbb{E}[Z_1], p_0^2 \cdot \mathbb{V}(Z_0) + p_1^2 \cdot \mathbb{V}(Z_1)).$$

We then sample new observations until we are at least 95% confident that $Z_D \neq 0$. In other words, when $p(Z_D < 0) < 0.025$, we accept $h = 0$, and when $p(Z_D < 0) > 0.975$, we accept $h = 1$. Because we compare the hypotheses in order to determine when to stop sampling, biasing the prior should result in requiring less evidence for one hypothesis before stopping, and more evidence for the other hypothesis.

To choose where to sample, we compute the expected variance of Z_h given a new observation at θ_a. From Osborne et al. (2012), we compute for each of the actions in Equation 5.4 the following:

$$\mathbb{E}[\mathbb{V}(Z_h|\theta_a)] = \mathbb{V}(Z_h) + \mathbb{E}[Z_h] - \int \mathbb{E}[Z_h|\theta_a]^2 \mathcal{N}(\mu_h(\theta_a), \mathrm{Cov}_h(\theta_a, \theta_a))\, \mathrm{d}\log S_h(\theta_a).$$

We then pick θ_a as:

$$\theta_a^* = \arg\min_{\theta_a} \mathbb{E}[\mathbb{V}(Z_h|\theta_a)].$$

5.2 Experiment 5.1: Mental Rotation of 2D Stimuli

To evaluate the models described previously, we ran a behavioral experiment based on classic mental rotation studies (e.g. Cooper, 1975; Shepard & Metzler, 1971).

5.2.1 Stimuli

We randomly generated 20 shapes of five or six vertices (e.g., Figure 5.1). For each shape, we computed 20 "same" and 20 "flipped" stimuli pairs, with 18 rotations (θ) spaced at $20°$ increments between $0°$ and $360°$ (with $0°$ and $180°$ repeated twice, in order to gather an equal number of responses for each angle between $0°$ and $180°$). "Same" pairs were created by rotating $\mathbf{X}_a$ by θ; "flipped" pairs were first reflected $\mathbf{X}_a$ across the y-axis, then rotated by θ.

We generated five additional shapes to be used in a practice block of 10 trials. Across these trials, there was one "flipped" and one "same" repetition of each shape and each angle ($60°$, $120°$, $180°$, $240°$, or $300°$) such that no shape was presented at the same angle twice. We also generated a sixth shape to include with the instructions. This shape had both a "flipped" and "same" version, each rotated to $320°$.

5.2.2 Participants and Design

We recruited 247 participants on Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015). Each participant was paid $1.00 for 15 minutes of work, consisting of one block of 10 practice trials followed by two blocks of 100 randomly ordered experiment trials.

All participants saw the same 10 practice trials as described above. There were 720 unique experimental stimuli (20 shapes × 18 angles × 2 reflections), though because stimuli with rotations of 0° or 180° were repeated twice, there were 800 total experimental stimuli. These stimuli were split across eight conditions in the following manner: first, stimuli were split into four blocks of 200 trials. Within each block, each shape was repeated ten times and each rotation was repeated ten times (five "same", five "flipped"), such that across all blocks, each stimulus appeared once. Each block was then split in half, and participants completed two half-blocks.

5.2.3 Procedure

Participants were given the following instructions while being shown an example "same" pair and an example "flipped" pair: *"On each trial, you will see two images. Sometimes, they show the* ***same*** *object. Other times, the images show* ***flipped*** *objects. The task is to determine whether the two images show the* ***same*** *object or* ***flipped*** *objects."*

On each trial, participants were instructed to press the 'b' key to begin and to focus on the fixation cross that appeared for 750ms afterwards. The two images were then presented side-by-side, each at 300px × 300px, and participants could press 's' to indicate they thought the images depicted the "same" object, or 'd' to indicate they thought the images depicted "flipped" objects. While there was no limit on RT, we urged participants to answer as quickly as possible while maintaining at least 85% accuracy in the experimental blocks.

5.2.4 Analysis

For analyses of RT, confidence intervals around harmonic means of correct responses were computed using a bootstrap analysis of 10000 bootstrap samples (sampled with replacement). We also used a bootstrap analysis of 10000 bootstrap samples to compute the confidence intervals around both Spearman (ρ) and Pearson (r) correlations. Unless otherwise specified, all correlations were computed over 720 stimuli. For analyses of accuracy, confidence intervals were computed from a binomial proportion with a Jeffrey's beta prior. To test if judgments were

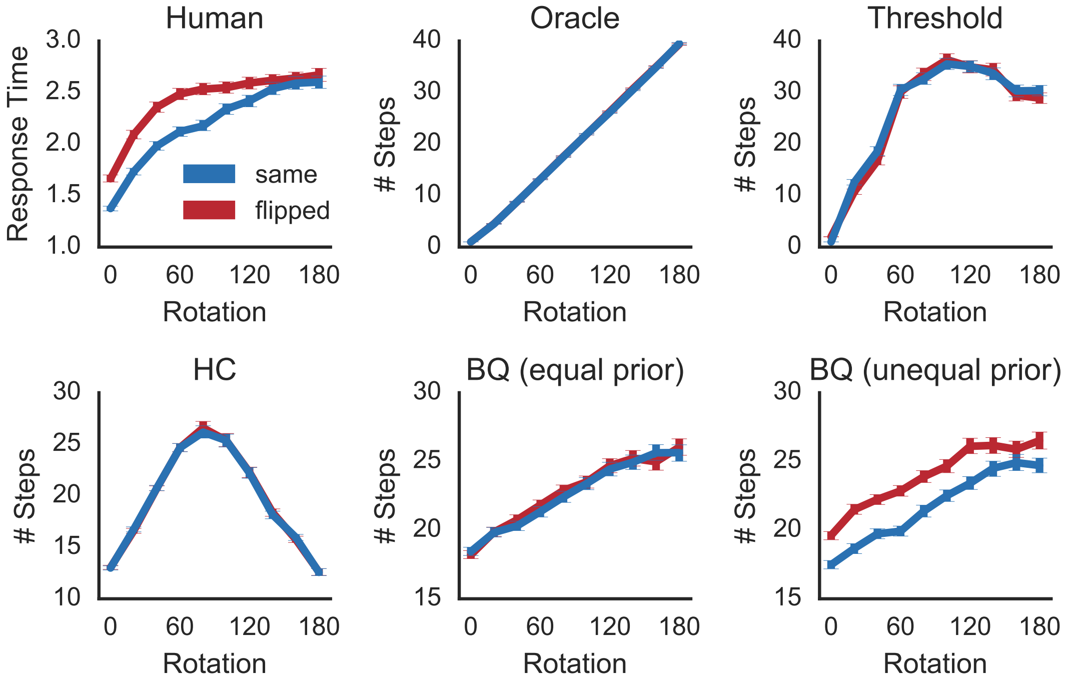


Figure 5.2: Response time comparison. RT of correct responses as a function of the minimum angle of rotation. All error bars are 95% confidence intervals.

above chance on a particular stimulus, we used the same binomial proportion and tested whether $p\,(p(\text{correct}) \leq 0.5) \leq \frac{0.05}{720}$, where $\frac{1}{720}$ is a Bonferroni correction for multiple comparisons.

5.2.5 Results

Of the 247 participants, 200 (81%) were included in our analyses. Of the other 47, we excluded 10 (4%) because of an experimental error, 6 (2.4%) because they had already completed a related experiment, and 31 (12.6%) because they failed a comprehension check, which was defined as correctly answering at least 85% of stimuli with a rotation of either $0°$, $20°$, or $340°$. We also excluded 82 trials for which the RT was either less than 100ms or greater than 20s.

The average RT across all correctly-judged stimuli was $M = 2248.6$ msec, 95% CI [2260.8 msec, 2236.9 msec]; the full histogram of RTs can be seen in Figure 5.4. The minimum angle of rotation was significantly rank-order (Spearman) correlated with average per-stimulus

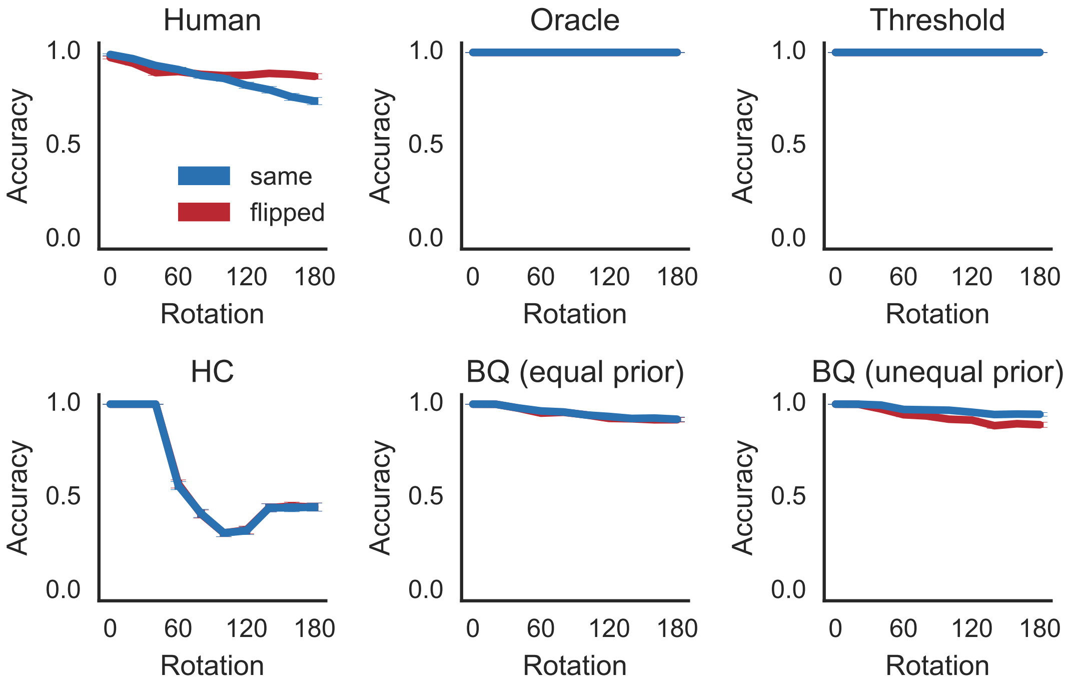


Figure 5.3: Accuracy comparison. Accuracy as a function of the minimum angle of rotation. All error bars are 95% confidence intervals.

RTs, both for "flipped" ($\rho = 0.50$, 95% CI $[0.41, 0.58]$) and "same" pairs ($\rho = 0.68$, 95% CI $[0.61, 0.73]$). While this replicates the general result of previous experiments (e.g. Cooper, 1975; Shepard & Metzler, 1971), our results are not as linear (Figure 5.2).

The average accuracy across all stimuli was $M = 88.2\%$, 95% CI $[87.9\%, 88.6\%]$, though there were 62 stimuli (out of 720) for which people were not above chance. The minimum angle was also correlated with participants' average per-stimulus accuracy, though much more so for "same" pairs ($\rho = -0.77$, 95% CI $[-0.81, -0.72]$) than "flipped" pairs ($\rho = -0.36$, 95% CI $[-0.46, -0.26]$). This is the same result found both by Cooper (1975) and Gardony et al. (2014).

There was a significant effect of trial number both on RT ($\rho = -0.76$, 95% CI $[-0.81, -0.69]$) and on accuracy ($\rho = 0.66$, 95% CI $[0.57, 0.73]$), though the effect on accuracy was not significant during the second half of the experiment ($\rho = 0.52$, 95% CI $[0.36, 0.66]$ for the first half vs. $\rho = 0.16$, 95% CI $[-0.02, 0.34]$ for the second half). These effects may have contributed to the not-quite-linearity of the human RTs; future work should collect more data per participant.

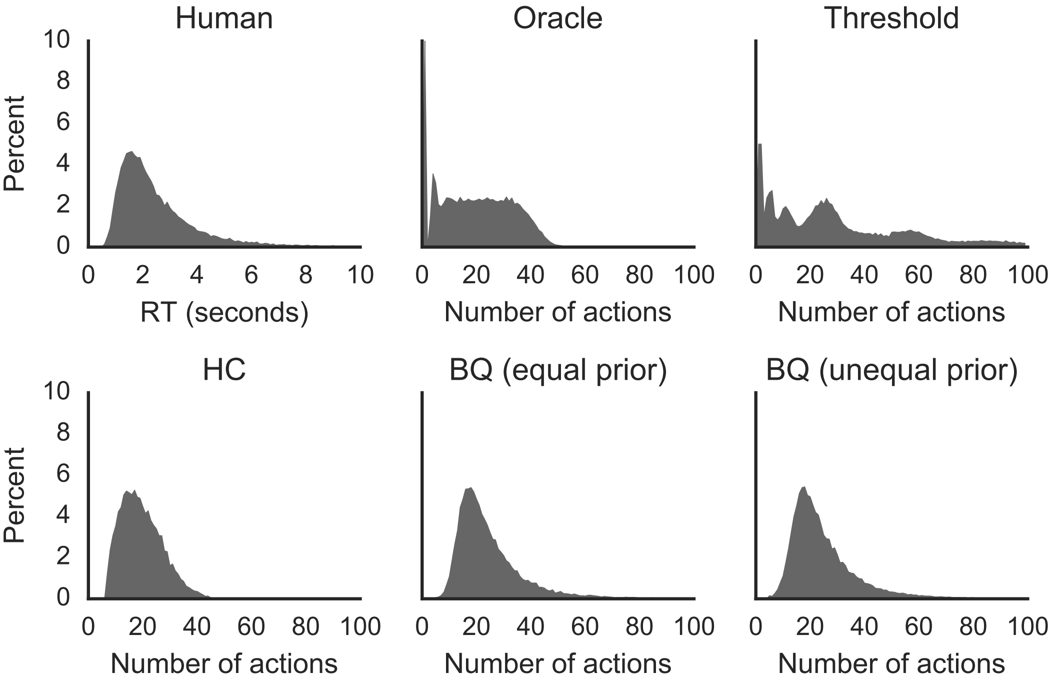


Figure 5.4: Response time histograms. Each subplot shows the distribution of RTs on correct trials for people and the models.

5.3 Comparing Human and Model Judgments

For each model, we ran 50 samples for each of the 800 experimental stimuli. The step size parameter (σ_ϵ) was fit to human RTs for each of the models, resulting in $\sigma_\epsilon = 0.6$ for the Threshold and BQ models and $\sigma_\epsilon = 0.1$ for the Oracle and HC models. We also ran the models under two different priors, $p(h = 0) = 0.5$ (the "equal" prior) and $p(h = 0) = 0.55$ (the "unequal" prior). As expected, this only had a major effect on the stopping criteria for the BQ model.

5.3.1 Oracle model

The number of actions taken by the Oracle model was perfectly correlated with the minimum angle of rotation (Figure 5.2). The Oracle model was the best fit to human RTs, with a correlation of $r = 0.57$, 95% CI $[0.52, 0.61]$ (Figure 5.5), although the distribution of response times did not match that of people (Figure 5.4). Moreover, the Oracle model was 100% accurate, and therefore

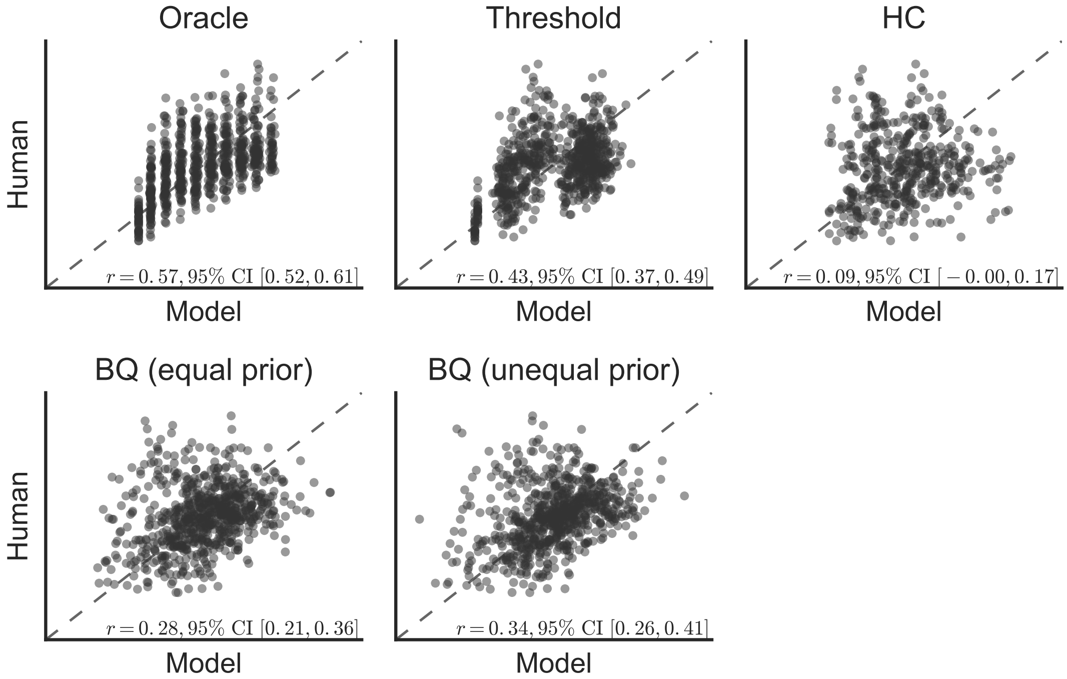


Figure 5.5: Model vs. human RTs. Each subplot shows the z-scored model RTs (x-axis) vs. the z-scored human RTs (y-axis). Pearson correlations are shown beneath each subplot. The dotted lines are $x = y$.

could not explain the effect of rotation on people's accuracy.

5.3.2 Threshold model

There was an overall monotonic relationship between the minimum angle of rotation and the number of actions taken by the Threshold model (Figure 5.2), though this relationship did not hold for *individual* shapes (e.g., Figure 5.6). The Threshold model was able to explain a moderate amount of the variance in human RTs, with a correlation of $r = 0.43$, 95% CI $[0.37, 0.49]$ (Figure 5.5). Like the Oracle model, the overall distribution of its RTs did not match that of people (Figure 5.4). The Threshold model had 100% accuracy, and thus did not exhibit a relationship between minimum angle and accuracy. As noted, we fit $\sigma_\epsilon = 0.6$ for the Threshold model. This had the interesting effect of causing the Threshold model to *over*rotate, because the step size was large enough that it sometimes missed the global maximum, and had to do another full rotation to find it.

5.3.3 HC model

The HC was the only model for which there was no monotonic relationship between rotation and RT (Figure 5.2). Moreover, the HC model was barely above chance ($M = 59.7\%$, 95% CI $[59.2\%, 60.2\%]$) and there were 312 stimuli for which it was not above chance. The HC model was not a good predictor of human RTs ($r = 0.09$, 95% CI $[-0.00, 0.17]$), as shown in Figure 5.5. It was a moderate predictor of human accuracy ($r = 0.24$, 95% CI $[0.17, 0.31]$).

5.3.4 BQ model

Like the Oracle and Threshold models, there was an overall monotonic relationship between rotation and the number of steps taken by the BQ model (Figure 5.2). Unlike the Threshold model, this relationship existed for individual shapes as well (e.g., Figure 5.6). The BQ model explained variance in human RTs about as well as the Threshold model (Figure 5.5), with a correlation of $r = 0.28$, 95% CI $[0.21, 0.36]$ for the equal prior and $r = 0.34$, 95% CI $[0.26, 0.41]$ for the unequal prior, and the RT distribution from the BQ model had the same overall shape as that of people (Figure 5.4).

The BQ model was quite accurate overall (equal prior: ($M = 95.3\%$, 95% CI $[95.1\%, 95.5\%]$; unequal prior: $M = 95.3\%$, 95% CI $[95.1\%, 95.5\%]$). With the equal prior, there were 12 stimuli for which it was not above chance; with the unequal prior, there were 14. The correlation with people's accuracy was $r = 0.23$, 95% CI $[0.16, 0.30]$ (equal prior) and $r = 0.15$, 95% CI $[0.08, 0.21]$ (unequal prior).

Because the BQ model relies on Equation 5.2 for its stopping criteria (as opposed to just finding a maximum), the prior $p(h)$ had an observable effect (Figure 5.2). As expected, with just a small bias of $p(h = 0) = 0.55$, there was a clear separation in RTs for "same" versus "flipped" stimuli: because of this bias, the model needed less evidence before accepting $h = 0$ (thus taking less time). This separation is similar to the trend also observed in human RTs. The prior also had an effect on accuracy (though this did not reflect human behavior): the bias towards $h = 0$ meant that the model was more likely to judge a pair as "same", thus, accuracy increased for "same" pairs, but decreased for "flipped" pairs.

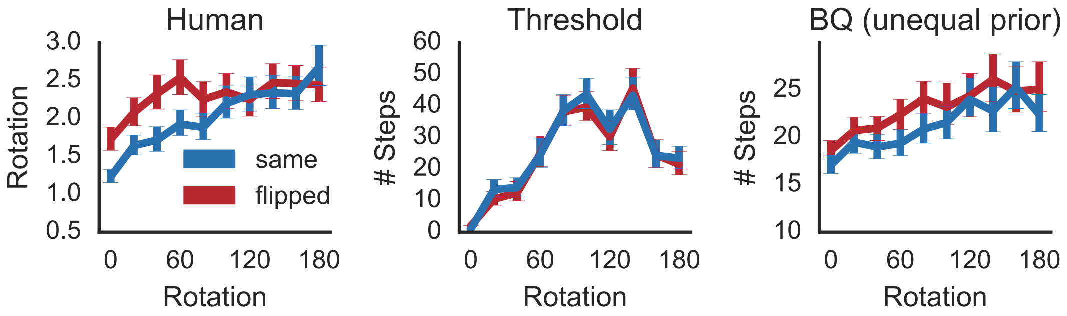


Figure 5.6: Typical RT curves for a single object. These plots correspond to the object shown in Figure 5.1. Left: human curves are either linear (as with the ``same'' pairs), or linear and then flat (as with the ``flipped'' pairs). Middle: the Threshold model does not have a monotonic relationship with rotation. Right: the BQ model is roughly linear.

5.4 General Discussion

We set out to answer the question of how people decide *what* to simulate when using mental imagery. Focusing on the specific case of determining the direction and extent of mental rotation, we formalized four models and compared their performance with the results of a behavioral experiment.

The Oracle and Threshold models were the best predictors of human RTs. However, both are somewhat unsatisfying explanations because they rely on *a priori* knowledge that people are unlikely to have. Moreover, they offer no explanation of several aspects of human behavior. First, their overall RT distributions look nothing like people's (Figure 5.4). Second, they both are 100% accurate, and so cannot explain the systematic relationship between rotation and human accuracy (Figure 5.3). Third, neither model can explain the difference in people's behavior on "same" and "flipped" stimuli.

In contrast, the BQ model was nearly as good as the Threshold model, yet it makes no assumptions about people's *a priori* knowledge. Furthermore, the BQ model matches people's behavior better than the Oracle or Threshold models in several ways. Its overall RT histogram has the same general shape as people's (Figure 5.4). Moreover, a closer look shows that the BQ model maintains the monotonic relationship between angle and RT even on individual stimuli, while the Threshold model does not (Figure 5.6). Finally, the BQ model's adaptive stopping rule is sensitive to the prior, and thus provides a possible explanation for why people are slower to respond on “flipped” stimulus pairs.

Thus, we suggest that the BQ model offers the most promising explanation of people’s behavior on the mental rotation task to date. While it is not a perfect account, there are several ways in which it could be improved. For example, while we used holistic rotations in this paper, there is evidence that people compare individual features of shapes (Just & Carpenter, 1976; Yuille & Steiger, 1982). Additionally, a different active sampling approach could maintain a distribution over the location and value of the global maximum, rather than over the integral. We intend to explore these possibilities in future work, building upon the foundation established in this paper and working towards a better understanding of *what* people choose to simulate.

Because when you are imagining, you might as well imagine something worth while.

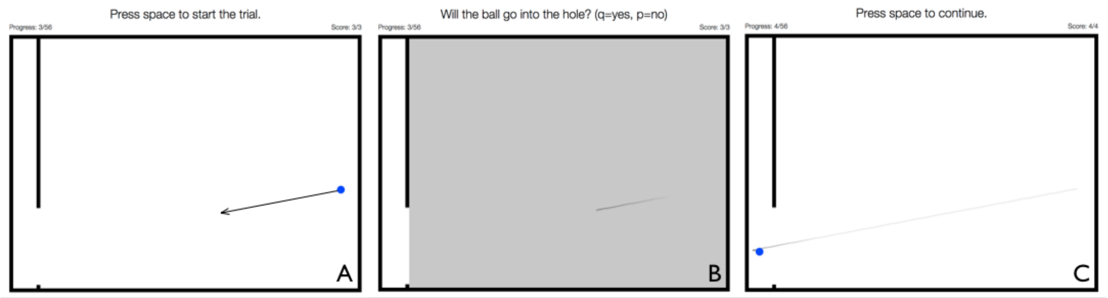
L.M. Montgomery, *Anne of Green Gables*

6 Allocation of Cognitive Resources

Consider the game of Angry Birds, where the goal is to launch birds to knock down a tower. To take a shot, the player can imagine—or *mentally simulate*—the path the bird will take and how it will affect the tower. How long should the player spend thinking before they let each bird fly? If they spend very little time thinking, they are likely to miss the target. But, if they spend too long thinking, it will take much longer to receive the satisfaction of beating the level. More generally, if running simulations will provide a more accurate forecast but incur a sampling cost, how long should an agent spend simulating before acting?

In the domain of physical reasoning, research suggests that people make predictions about physical scenes—such as those found in Angry Birds—by running noisy physical simulations (Battaglia et al., 2013; Gerstenberg et al., 2014; Sanborn et al., 2013; Smith, Battaglia, & Vul, 2013; Smith, Dechter, et al., 2013; Smith & Vul, 2013, 2014; Ullman et al., 2014). However, while this research has investigated the *mechanism* for making these predictions, there has been very little investigation into how people *use* this mechanism. In particular, because the simulations are noisy, it may be beneficial to run multiple simulations in order to obtain more accurate predictions. Is there an optimal number of simulations to run in these situations? If so, do people behave optimally?

Experiment 6.1 of this chapter was previously published in Hamrick, Smith, Griffiths, and Vul (2015).



Press space to start the trial.

Progress: 3/56

Score: 3/3

Will the ball go into the hole? (q=yes, p=no)

Progress: 3/56

Score: 3/3

Press space to continue.

Progress: 4/56

Score: 4/4

Figure 6.1: Example experimental trial. Each panel shows a different part of the trial. *A:* the initial screen presented to the participant. The arrow was not part of the actual stimuli; it has been added to reflect the animation that participants observed after pressing ``space''. *B:* the screen is occluded after observing the stimulus presentation. The faded gray line shows the path the ball took during the initial presentation. *C:* the final position of the ball, after observing the feedback. As in the middle panel, the faded gray line shows the path of the ball.

To investigate how many simulations people run, we focus on a dichotomous prediction task—will a ball in motion on a computer screen go through a hole, or miss it? To model this task, we combine a mechanism of noisy physical simulation (Smith & Vul, 2013) with a decision strategy for sample-based agents known as the *sequential probability ratio test*, or the SPRT (Wald, 1947). An agent acting according to this strategy takes samples that point to one hypothesis or another, and continues to do so until the net samples in favor of one hypothesis reaches a threshold, at which point that hypothesis wins. Following the approach of rational analysis (Anderson, 1990), we choose the SPRT as our model of choice because it provides an optimal cost-benefit trade-off between sampling and exploiting information (Wald & Wolfowitz, 1950). Additionally, the SPRT's continuous analogue—the drift-diffusion model—has been widely used to explain behavior in a number of decision-making tasks (e.g. Bitzer, Park, Blankenburg, & Kiebel, 2014; Gold & Shadlen, 2007; Ratcliff & McKoon, 2008), lending further credibility to the SPRT as a model of human cognition.

The SPRT strategy predicts that people need to take more samples—and thus also will take a longer time to respond—when there is roughly equal evidence for both hypotheses. Additionally, a meta-level analysis of the SPRT predicts that when the payoff structure in the world changes (for example, when time is of the essence) people should adjust their evidence accumulation threshold to reflect the resulting speed/accuracy trade-off (Vul et al., 2014). Based on these predictions, we hypothesize that people make decisions by running mental simulations, and that they vary the number of simulations based both on their uncertainty as well as the cost structure of the world. In the first two sections of this chapter, we test these predictions through three experiments in which we asked participants to respond to the question of, "will the ball go through the hole?", and analyze peoples' judgments and response times. Next, we formalize our model, combining the simulation model from Smith and Vul (2013) with the SPRT decision strategy. We then demonstrate that our model can explain the empirical pattern of responses and response times within each experiment. We also show that, across experiments, people adjust not only their evidence accumulation threshold, but also the detail of the simulations themselves. Finally, we discuss the implications of our results on the broader, underlying question: how should people make use of mental simulations?

6.1 Experiment 6.1: Testing the Effect of Uncertainty

To determine whether people adapt the amount of time they spend to make a decision, we ran an experiment in which people made a binary judgment about whether a ball traveling across a computer screen would go through a hole (see Figure 6.1). We designed the trials to elicit a range of responses by varying the margin by which the ball either missed or went through the hole. According to adaptive decision making models like the SPRT, when people's simulations are uncertain—i.e., when the probability that the ball goes in the hole is close to $p = 0.5$, such as when the ball just barely goes through the hole—they should be slower to respond. People should be faster to respond when their simulations are more certain, such as when the ball misses the hole by a wide margin.

6.1.1 Participants

We recruited 328 participants on Amazon's Mechanical Turk using the psiTurk (Gureckis et al., 2015) experimental framework. Participants were treated in accordance with UC Berkeley IRB standards and were paid $0.60 for 6.5 minutes of work. Participants were randomly assigned to one of eight conditions, which determined which stimuli they judged (see Stimuli). We excluded 8 participants for answering incorrectly on more than one control trial (see Stimuli), leaving a total of 320 participants.

6.1.2 Procedure

On each trial, participants were shown the scene, including the initial position of the ball and the location of the hole. Participants were instructed to press "space" to begin the trial, after which an animation of the initial stimulus began (see Stimuli). As soon as this animation concluded, a gray box was drawn over the screen, occluding the ball (but not the line depicting the path it had traveled so far; this was left in as a reminder of where the ball had come from). Participants were asked, "will the ball go in the hole?", and were instructed to press 'q' if they thought it would, and 'p' otherwise. After responding, text appeared saying "Correct!" or "Incorrect." The gray occluder was removed, and participants were shown a feedback animation of the path of the ball (see Stimuli). The final frame of this animation remained on the screen until participants pressed "space" to advance to the next trial.

Participants were given seven instruction trials prior to the experiment to familiarize them with the procedure. Then, participants made judgments on 48 experimental trials in a random order. There were also eight control trials, which were shown in a random order after every seven experiment trials.

6.1.3 Stimuli

The stimuli consisted of two animations—the *stimulus presentation* and the *feedback* animations—depicting a blue ball with a radius of 10px moving in a box with dimensions 900px × 650px. In both animations, the ball had a velocity of 400px/s, and as it moved, it traced a gray line (see Figure 6.1). The stimulus presentation had a duration of 0.75s and depicted the ball moving in a particular direction. The feedback had a duration of 1.25s and picked up where the stimulus presentation left off; it showed the ball either going into the hole or bouncing off the wall that contained the hole. Across all stimuli, the ball traveled the same distance during both animations, and could bounce on the other walls 0, 1, or 2 times before going into the hole or hitting the wall with the hole.

There were 48 different initial animations, equally balanced by number of bounces during feedback (16 each for 0, 1, and 2 bounces). For each of these initial animations, there were four trial types and two hole sizes, for a total of eight versions of each stimulus. The four trial types were: "far in" (FI), where the ball went through the center of the hole; "far miss" (FM), where the ball missed the hole by a wide margin; "close in" (CI), where the ball just barely went through the hole; and " close miss" (CM), where the ball just barely missed the hole. The two hole sizes were 100px and 200px.

In order to ensure that participants never saw the same initial animation twice, we used a Latin square design of Initial Animation × Trial Type × Hole Size. Thus, each participant saw each initial animation once, each trial type 12 times, and each hole size 24 times. This also ensured that the ball would go through the hole half the time, so that participants would not be biased

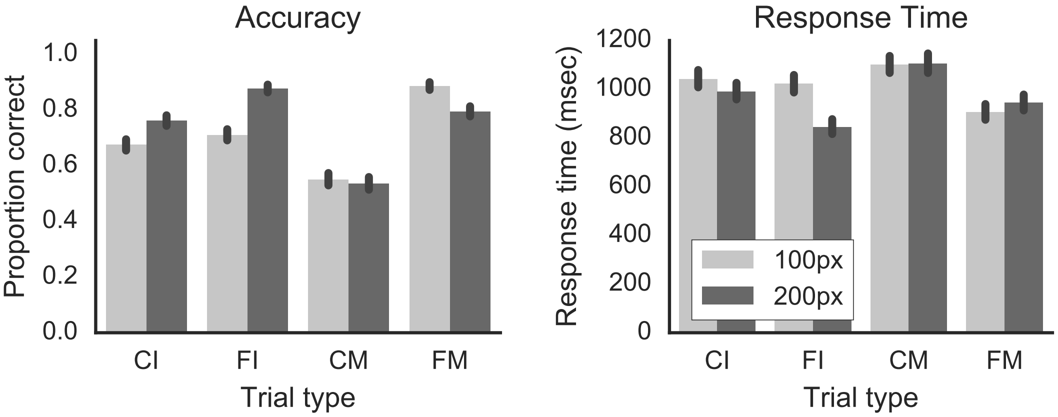


Figure 6.2: Responses and RTs as a function of trial type. Left: Each bar shows the proportion of participants saying that the ball will go in the hole for a particular trial type (x-axis) and hole size (color). Right: Like the left subplot, but the y-axis shows bootstrapped logarithmic means of RTs. See Section 6.1.3 for an explanation of the trial types.

to respond either way. Additionally, there were seven instruction trials and eight control trials, which were the same for all participants. The control trials were designed to be easy and were either of type "straight hit" (with a hole size of either 300px or 350px) or "far miss" (with a hole size of 100px). Thus, participants saw a total of 63 trials.

6.1.4 Results

6.1.4.1 Responses

On average, participants were correct $72.4\%,\ 95\%$ CI $[71.7\%, 73.1\%]$ of the time and responded that the ball would go in the hole $53.2\%,\ 95\%$ CI $[52.4\%, 54.0\%]$ of the time, indicating a slight bias towards believing the ball would go in. There was a significant effect of trial type on participants' responses ($\chi^2(3) = 524.245, p < 0.001$) as well as a significant effect of hole size ($\chi^2(1) = 39.220, p < 0.001$). There was also an interaction between trial type and hole size ($\chi^2(3) = 244.313, p < 0.001$). There was a significant difference between responses for the two hole sizes (for CI, $z = -6.26, p < 0.001$; for FI, $z = -13.08, p < 0.001$; and for FM, $z = 7.90, p < 0.001$) except on the CM trials ($z = 0.85, p = 0.39$). Figure 6.2 shows responses as a function of trial type and hole size.

6.1.4.2 *Response times*

For all analyses of response time (RT), we computed averages using bootstrapped logarithmic means (exponential of the mean of the log RTs), using 10000 bootstrap samples. On average, participants responded in $RT = 990.84\,\mathrm{msec},\ 95\%\ \mathrm{CI}\ [978.30, 1002.97]$, excluding catch trials. There were effects of both trial type ($\chi^2(3) = 102.356, p < 0.001$) and hole size ($\chi^2(1) = 6.603, p < 0.05$), as well as an interaction between trial type and hole size ($\chi^2(3) = 80.108, p < 0.001$). As in Figure 6.2, hole size only had an effect in the case of the CI ($z = 2.57, p < 0.05$) and FI trials ($z = 9.88, p < 0.001$). For CM trials, this difference was not significant ($z = -0.19, p = 0.85$) and for FM trials, it was only marginally significant ($z = -1.79, p = 0.07$).

Participants were fastest to respond on trials with zero bounces ($RT = 779.53\ \mathrm{msec},\ 95\%\ \mathrm{CI}\ [762.89, 796.55]$), slower to respond on trials with one bounce ($RT = 1015.17\ \mathrm{msec},\ 95\%\ \mathrm{CI}\ [994.88, 1036.48]$), and slowest to respond on trials with two bounces ($RT = 1228.65\ \mathrm{msec},\ 95\%\ \mathrm{CI}\ [1203.26, 1254.37]$).

6.1.4.3 *Relationship of responses and RTs*

According to the SPRT, participants should be slower on trials for which they are less certain (i.e., when their average response is closer to 0.5), and faster when they are more certain (i.e., when their average response is closer to 0 or 1). Figure 6.3 illustrates that this trend does indeed appear. To demonstrate this trend more quantitatively, we constructed a linear mixed effects model for average log-transformed response times as a function of hole width, trial type, and the number of bounces (as well as their interactions). We compared this model to a second one that additionally included a term for people's average responses. We also compared to a third model that also included a second-order response term. Both the first- and second-order terms were significant predictors of people's response times ($\chi^2(1) = 11.768, p < 0.001$ for the first-order term; $\chi^2(1) = 49.641, p < 0.001$ for the second-order term). These results confirm the hypothesis that people take longer to respond on trials where they are more uncertain, even when controlling for other factors that might influence response time.

6.1.4.4 *Learning*

To check for practice effects, we computed Spearman rank correlations (with 95% confidence intervals computed from 10000 bootstrap samples) between trial number and accuracy, as well as between trial number and RT. We found an overall effect of practice on accuracy ($\rho = 0.37,\ 95\%\ \mathrm{CI}\ [0.11, 0.44]$), though in the second half of the experiment, this effect

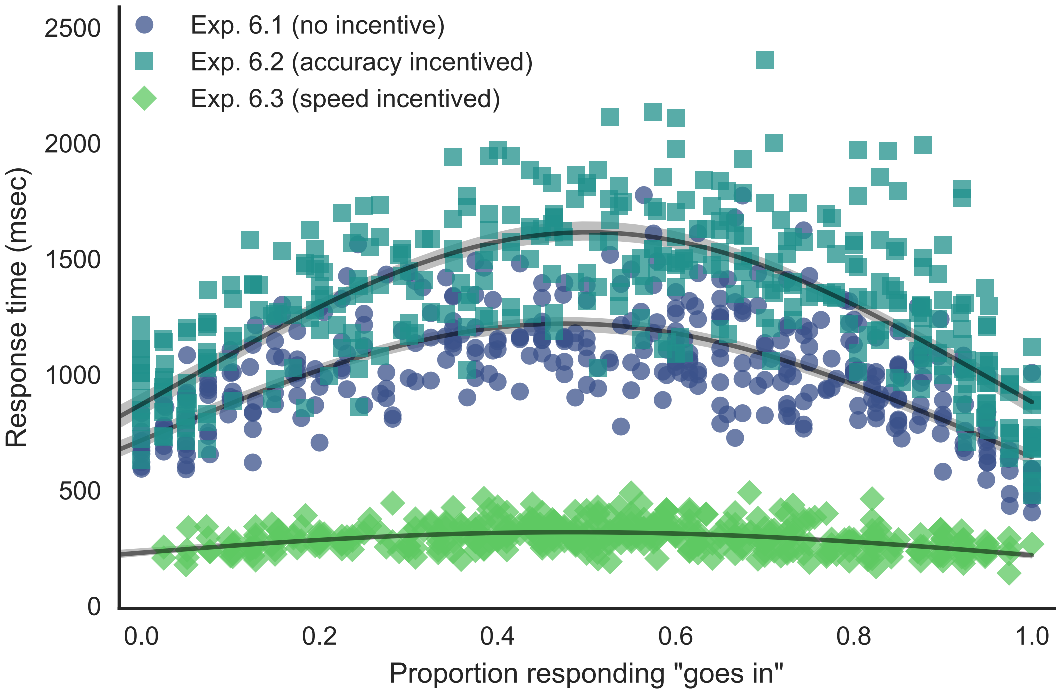


Figure 6.3: RTs as a function of responses. Each point corresponds to a separate stimulus, with measures averaged across participants. The x-axis shows average responses, while the y-axis shows log-averaged response times. The color and shape of the points are redundant, and reflect Experiments 6.1, 6.2, and 6.3. The black lines correspond to 2^{nd} order best fit lines, with shaded regions indicating 95% confidence intervals. As discussed in the text, the relationships between responses and RTs are the same between Experiments 6.1 and 6.2 (with the exception of a difference in intercept), while the relationships between Experiments 6.1 and 6.3 are significantly different, with the relationship in Experiment 6.3 being much less strong.

disappeared ($\rho = -0.07,\ 95\%\ \mathrm{CI}\ [-0.37, 0.19]$). There was also an overall effect of practice on RT ($\rho = -0.96,\ 95\%\ \mathrm{CI}\ [-0.93, -0.85]$), which was strong both in the first ($\rho = -0.94,\ 95\%\ \mathrm{CI}\ [-0.95, -0.83]$) and second ($\rho = -0.86,\ 95\%\ \mathrm{CI}\ [-0.81, -0.37]$) halves of the experiment. These results are shown in Figure 6.4.

6.1.5 Discussion

The results from Experiment 6.1 indicate that people change the amount of time they spend deciding whether the ball will go through the hole in a way that is systematically related to how

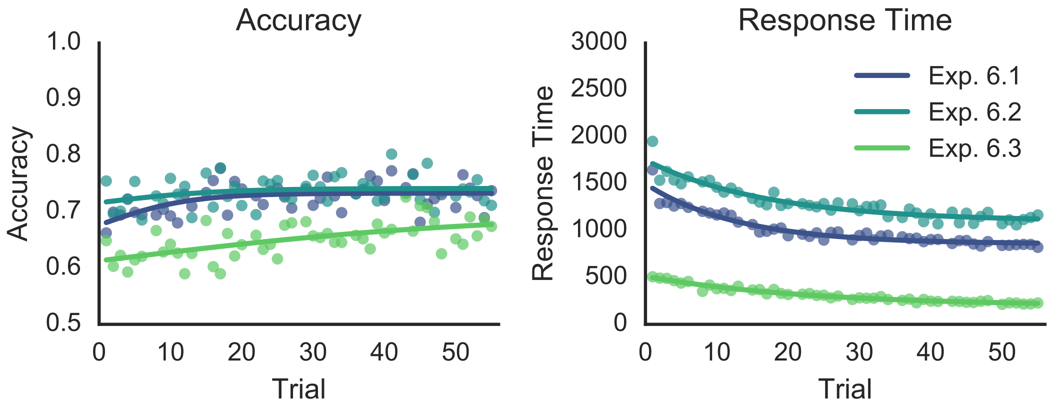


Figure 6.4: Effect of practice. *Left*: average accuracy as a function of trial. Points indicate averages for each trial, and the lines are best-fit sigmoid functions. *Right*: log-averaged response times as a function of trial. Points indicate averages for each trial, and the lines are best-fit exponential functions.

likely they think the ball is to go in. In other words, people take longer to respond when they are more uncertain, and less time to respond when they are more certain—a result that is qualitatively predicted by the SPRT. We will go into further detail analyzing peoples' responses and response times with respect to the SPRT in Section 6.3.

6.2 Experiments 6.2-6.3: Raising and Lowering the Stakes

The results of Experiment 6.1 raise a follow-up question: to what extent do people change the amount of time they spend thinking about a problem when the payoff structure of the world changes? To answer this question, we ran two additional experiments that were identical to Experiment 6.1 except that the payoff structure was different. In Experiment 6.2, we incentivized people to respond more accurately, while in Experiment 6.3, we incentivized people to respond more quickly.

6.2.1 Participants

We recruited 327 (Experiment 6.2) and 395 (Experiment 6.3) participants on Amazon's Mechanical Turk using the psiTurk (Gureckis et al., 2015) experimental framework. Participants were treated in accordance with UC Berkeley IRB standards and were paid for approximately 7.25

minutes of work in Experiment 6.2 and 6 minutes of work in Experiment 6.3. In Experiment 6.2, we paid people a base pay of $0.35 as well as an average bonus of $0.29 that was based on the number of correct trials: if they had an overall accuracy less than 70%, they received no bonus; if they had an accuracy between 70% and 80%, they were paid $0.005 for each correct answer; if they had an accuracy between 80% and 90%, they were paid $0.01 for each correct answer; and if they had an accuracy greater than 90%, they were paid $0.015 for each correct answer. In Experiment 6.3, we paid people a base pay of $0.20 as well as an average bonus of $0.38 that was based on the speed they responded: if they responded within 100msec, they were paid a bonus of $0.01, which decreased linearly with time up to 1000msec, at which point they would receive no bonus.

As in Experiment 6.1, participants were randomly assigned to one of eight conditions, which determined which stimuli they judged (see Stimuli). In Experiment 6.2, we excluded 5 participants for answering incorrectly on more than one control trial, leaving a total of 322 participants. In Experiment 6.3, we excluded 76 participants for answering incorrectly on more than one control trial, leaving a total of 319 participants.

6.2.2 Procedure and Stimuli

The overall procedure and stimuli were the same as in Experiment 6.1, except that participants saw a counter indicating how much of a bonus they would receive. In Experiment 6.2, this was an estimated bonus (as the final bonus depended on overall accuracy). In Experiment 6.3, this was the actual bonus; additionally, during the time that participants could respond, a circular timer was displayed on the screen indicating the proportion of the maximum bonus that could be achieved.

6.2.3 Results

6.2.3.1 Responses

Participants responded correctly on $73.6\%,\ 95\%$ CI $[72.9\%, 74.3\%]$ of trials in Experiment 6.2, compared with $65.0\%,\ 95\%$ CI $[64.3\%, 65.8\%]$ in Experiment 6.3. We compared these accuracies to that from Experiment 6.1 ($72.4\%,\ 95\%$ CI $[71.7\%, 73.1\%]$), shown in Figure 6.5, by constructing a generalized linear model of accuracy as a function of experiment version, with contrasts corrected for multiple comparisons using Tukey's HSD. Between Experiments 6.1 and 6.2, there was only a slight, marginally significant difference in accuracy ($z = -2.34, p = 0.05$).

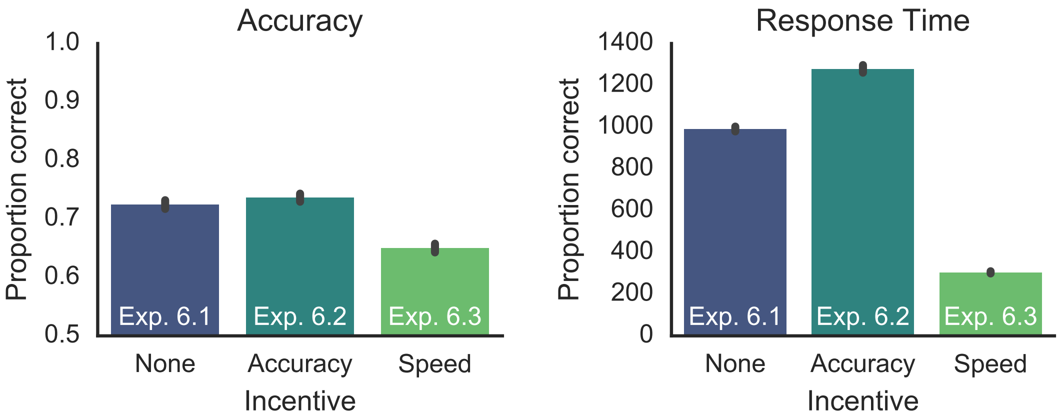


Figure 6.5: Average accuracy and response times. In both plots, the x-axis corresponds to the experiment and the y-axis corresponds to average accuracy (left) or log-average response times (right), with error bars reflecting bootstrapped 95% confidence intervals. Accuracy is nearly the same for Experiments 6.1 and 6.2, but is much lower in Experiment 6.3. Response times differ significantly between all three experiments.

Participants were significantly less accurate in Experiment 6.3 than they were in Experiment 6.1 ($z = 13.94, p < 0.001$).

6.2.3.2 *Response times*

Participants took on average $RT = 1277.88$ msec, 95% CI $[1260.48, 1295.55]$ to respond in Experiment 6.2, compared with $RT = 304.05$ msec, 95% CI $[298.70, 309.36]$ in Experiment 6.3. We compared these response times to those from Experiment 6.1 ($RT = 990.84$ msec, 95% CI $[978.30, 1002.97]$), shown in Figure 6.5, by constructing a linear model of log-transformed response times as a function of experiment version, with contrasts corrected for multiple comparisons using Tukey's HSD. People took significantly more time to respond in Experiment 6.2 than in Experiment 6.1 ($t(45751) = -23.75, p < 0.001$), and significantly less time to respond in Experiment 6.3 than in Experiment 6.1 ($t(45751) = 110.10, p < 0.001$).

6.2.3.3 *Relationship of responses and RTs*

We found significant effects of second-order relationships between people's responses and response times in both Experiments 6.2 ($\chi^2(1) = 66.007, p < 0.001$) and 6.3 ($\chi^2(1) =$

$15.771, p < 0.001$), following the same analysis reported for Experiment 6.1. To compare whether these relationships differed between the experiments, we constructed a linear model for average log-transformed response times as a function of first- and second-order response terms, experiment version, and interaction terms between responses and versions. We then compared intercepts for different terms, relative to Experiment 6.1. We found significant non-zero intercepts for Experiment 6.2 ($\beta = 0.19,\ 95\%$ CI $[0.12, 0.26]$; $t(1143) = 5.49, p < 0.001$) and Experiment 6.3 ($\beta = -1.10,\ 95\%$ CI $[-1.19, -1.01]$; $t(1143) = -24.36, p < 0.001$), reflecting the differences in overall response times between the experiments. We did not find a significant first-order response term for Experiment 6.2 ($\beta = 0.23,\ 95\%$ CI $[-0.09, 0.54]$; $t(1143) = 1.41, p = 0.16$), though we did for Experiment 6.3 ($\beta = -0.91,\ 95\%$ CI $[-1.29, -0.53]$; $t(1143) = -4.75, p < 0.001$). Similarly, we did not find a significant second-order response term for Experiment 6.2 ($\beta = -0.11,\ 95\%$ CI $[-0.41, 0.19]$; $t(1143) = -0.71, p = 0.48$), though we did for Experiment 6.3 ($\beta = 0.97,\ 95\%$ CI $[0.62, 1.33]$; $t(1143) = 5.36, p < 0.001$). What these results suggest is that there was no difference in the response-vs-RT relationships between Experiments 6.1 and 6.2, apart from a difference in intercept (i.e., people in Experiment 6.2 took longer to respond). However, there was a difference in this relationship between Experiments 6.1 and 6.3—in particular, that this relationship was not as strong in Experiment 6.3.

6.2.3.4 *Learning*

To check for practice effects, we computed Spearman rank correlations (with 95% confidence intervals computed from 10000 bootstrap samples) between trial number and accuracy, as well as between trial number and RT. In Experiment 6.2, we found no overall effect of practice on accuracy ($\rho = 0.15,\ 95\%$ CI $[-0.03, 0.30]$), though there was an effect on response time ($\rho = -0.93,\ 95\%$ CI $[-0.91, -0.81]$) which persisted through both halves of the experiment. In Experiment 6.3, there was an overall effect of practice on accuracy ($\rho = 0.60,\ 95\%$ CI $[0.32, 0.61]$), though this effect disappeared during the second half of the experiment ($\rho = 0.24,\ 95\%$ CI $[-0.19, 0.47]$). There was also a strong effect of practice on response time ($\rho = -0.97,\ 95\%$ CI $[-0.96, -0.91]$). These results are shown in Figure 6.4.

6.2.4 Discussion

The results from Experiments 6.2 and 6.3 indicate that people are sensitive to the payoff structure in the task of predicting whether a ball will go through a hole, and that they change their behavior based on this. From the results of Experiment 6.2 (in which we incentivized people to be more accurate), we found that people do spend more time thinking about the task—however, that this doesn't lead to a significant increase in accuracy. Rather, it seems that the increased emphasis on accuracy led people to learn more quickly, as evidenced by a lack of practice effects on accuracy as compared to Experiment 6.1 (see also Figure 6.4). In contrast, we found that participants in Experiment 6.3 both decreased their thinking time (and correspondingly, their accuracy) in such a way that they spent nearly a constant amount of time thinking about each stimulus.

6.3 Modeling Differences in Response Times

The results from the previous experiments provide a compelling demonstration that there is a nonlinear relationship between how certain people are when judging a particular stimulus, and how much time they spend thinking about it. Moreover, as the results of Experiment 6.2 show, this relationship changes when people are put under time pressure. To model these results more quantitatively, we constructed a model of responses and response times based on a model of noisy physical simulation (Smith & Vul, 2013) combined with the *sequential probability ratio test*, or the SPRT (Wald, 1947).

6.3.1 Resource-Rational Analysis

6.3.1.1 Computational-level analysis

We begin with our formalization by performing a resource-rational analysis of the problem that people are solving (Griffiths et al., 2015). Specifically, to answer the question of "will the ball go in the hole?", the following equation must be computed:

$$\mathbf{X}_{0:T} = \pi(\mathbf{x}_0, \mathbf{v}_0, S), \tag{6.1}$$

where π is a function of physical dynamics, $\mathbf{x}_0$ is the initial position of the ball, $\mathbf{v}_0$ is the initial velocity of the ball, $\mathbf{X}_{0:T}$ is the path of the ball, and S is the rest of the scene (e.g. the walls, the location of the hole, etc.). Once $\mathbf{X}_{0:T}$ is computed, it is straightforward to determine whether the path of the ball passed through the hole: $\exists\ t_i, t_j \in [0, T)$ s.t. $(x_{t_i} > w_x) \wedge (x_{t_j} < w_x)$, where x_{t_i} is the ball's position along the x-axis at time t_i, and w_x is the wall's position along the x-axis.

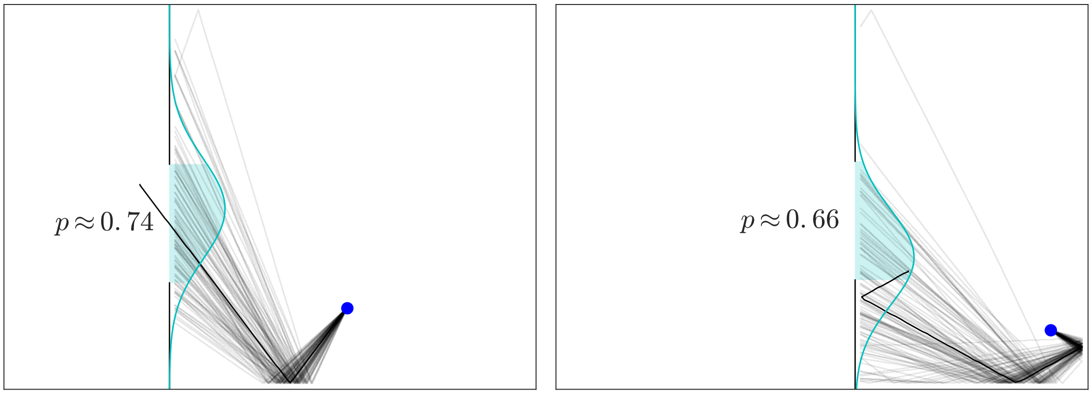


Figure 6.6: Illustration of simulation model. The images show 100 simulations from the physical model for two separate trials, with the left image depicting a ``far in'' trial with the large hole size, and the right image depicting a ``close miss'' trial with the large hole size. In both cases, the thick black lines indicate the ball's true trajectory, while the thin gray lines are simulations. The cyan lines indicate the distributions over the ball's endpoints, and the shaded regions show the section of the distributions overlapping the hole. The reported p-value is the probability under the distributions that the ball goes through the hole.

6.3.1.2 *Algorithmic approximation*

In the general case, the process of physical dynamics π may not be exactly known, or may be extremely difficult to compute exactly. To reflect uncertainty in the dynamics (which may arise either through approximation noise, or perceptual uncertainty), we write the dynamics as a probability distribution:

$$\mathbf{X}_{0:T} \sim \widetilde{\pi}(\mathbf{x}_0, \mathbf{v}_0, S; \sigma_p, \kappa_v, \kappa_m, \kappa_b), \tag{6.2}$$

where $\widetilde{\pi}$ is a noisy approximation to the true dynamics, parameterized by perceptual uncertainty of the ball's location (σ_p) and velocity (κ_v), dynamics uncertainty κ_m, and bounce angle uncertainty κ_b. For further details on the parameters of this formulation, see Smith and Vul (2013).

Given N samples $\{\mathbf{X}_{0:T}^{(n)}\}_{n=1}^{N}$ from the model in Equation 6.6, we can estimate the probability that the ball will go in the hole by looking at the distribution of endpoints. First, we model the distribution of endpoints as a normal distribution, with parameters $\hat{\mu}_{\mathrm{sim}}$ and $\hat{\sigma}_{\mathrm{sim}}$:

$$\hat{\mu}_{\mathrm{sim}} = \frac{1}{N}\sum_{n=1}^{N} y_T^{(n)}, \qquad \hat{\sigma}^2_{\mathrm{sim}} = \frac{1}{N}\sum_{n=1}^{N}(y_T^{(n)} - \hat{\mu}_{\mathrm{sim}})^2.$$

Given these MLE parameter estimates, we also incorporate the "center bias" parameter from Smith and Vul (2013), corresponding to a Gaussian prior located at the center of the screen. This prior yields a posterior Gaussian distribution over endpoints of the ball:

$$\sigma_y^2 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\hat{\sigma}_{\mathrm{sim}}^2}\right)^{-1}, \qquad \mu_y = \left(\frac{y_{\mathrm{center}}}{\sigma_0^2} + \frac{\hat{\mu}_{\mathrm{sim}}}{\hat{\sigma}_{\mathrm{sim}}^2}\right) \cdot \sigma_y^2,$$

where y_{center} is the center of the box along the y-axis. With this distribution, we can integrate over the location of the hole to determine the probability of the ball going in:

$$\begin{aligned} Z &= \int_{y_{\mathrm{min}}}^{y_{\mathrm{max}}} \mathcal{N}(\mu_y, \sigma_y)\, \mathrm{d}y, \\ p &\approx \frac{1}{Z} \int_{h_l}^{h_u} \mathcal{N}(\mu_y, \sigma_y)\, \mathrm{d}y, \end{aligned} \tag{6.3}$$

where h_l and h_u are the lower and upper bounds of the hole, respectively; y_{min} and y_{max} are the lower and upper bounds of the box, respectively, and σ_0 is the center bias parameter. Figure 6.6 illustrates example trajectories sampled from this model and corresponding probability estimates computed from Equation 6.3.

6.3.1.3 *Optimizing for limited resources*

We can estimate the probability that the ball goes through the hole using Equation 6.3, but this has an additional free parameter: the number of samples, N. Based on the empirical results reported in Experiments 6.1-6.3, people's response time varied as a function of their uncertainty. This suggests that, if people are in fact using a sampling strategy to solve the problem, that they are running a variable number of samples rather than relying a constant amount of computation. To formalize this adaptivity, we used the *sequential probability ratio test*, or the SPRT (Wald, 1947), which is the optimal procedure for making a decision as quickly as possible while maintaining a fixed level of accuracy across all possible decisions. The SPRT procedure accumulates binary evidence $Y_N = \sum_{n=1}^{N} 2X_n - 1$ until a threshold T is reached ($|Y_N| = T$), corresponding to a random walk over the natural numbers initialized at 0 and with absorbing boundaries at T and $-T$.

Given the threshold T and the probability of observing a positive sample, p, the SPRT defines distributions over the number of samples N and the final response, r. First, the probability of a positive response (i.e., that the ball will go in the hole) can be derived from Feller (1968, Eq. 2.4,

p. 345) as:

$$p(r = 1 \mid T, p) = \frac{p^T}{p^T + (1-p)^T}. \tag{6.4}$$

Similarly, the probability of the number of samples can be derived from Feller (1968, Eq. 5.7, p. 353) as:

$$\begin{aligned} C_{N,T} &= \frac{2^N}{2T} \sum_{v=1}^{2T-1} \cos^{N-1}\left(\frac{\pi v}{2T}\right) \sin\left(\frac{\pi v}{2T}\right) \sin\left(\frac{\pi v}{2}\right) \\ p(N \mid T, p) &= C_{N,T} \cdot \left[p^{\frac{N-T}{2}}(1-p)^{\frac{N+T}{2}} + (1-p)^{\frac{N-T}{2}} p^{\frac{N+T}{2}}\right], \end{aligned} \tag{6.5}$$

where $N \geq T$ and N has the same parity as T. Derivations for Equations 6.4 and 6.5 are given in Appendices B.3.1 and B.3.2, respectively.

6.3.1.4 Modeling response times

Equation 6.3 gives us the probability p that the ball goes in the hole, which can then be used along with a threshold parameter T to compute the actual response (Equation 6.4) and number of samples (Equation 6.5) according to the SPRT. While we can evaluate people's responses directly under the distribution $p(r \mid T, p)$, we cannot directly compare numbers of samples and response times. Instead, we model response times as a function of the number of samples (N), the path length of the ball (ℓ), and the number of times the ball bounces (B):

$$p(t \mid N, B, \ell, t_d, t_B, t_\ell, \sigma_t) = \mathcal{N}(t; t_d + N \cdot (t_B \cdot B + t_\ell \cdot \ell), \sigma_t), \tag{6.6}$$

where σ_t is the response noise and t_d, t_B, and t_ℓ are parameters corresponding to non-decision time, the time it takes to resolve a bounce, and the time it takes to simulate the path of the ball, respectively. Importantly, we included both ℓ and B as factors in response time as these were both significant predictors of response time in a related experiment above and beyond that expected from uncertainty alone (see Appendix B.2).

We can now compute the distribution over response times by marginalizing out the number of samples, the number of bounces, and the path length:

$$p(t \mid T, p, t_d, t_B, t_\ell, \sigma_t) = \int \sum_{B=0}^{\infty} \sum_{N=T}^{\infty} p(t \mid N, B, \ell) p(N \mid T, p) p(B, \ell \mid \mathbf{x}_0, \mathbf{v}_0, S) \, \mathrm{d}\ell,$$

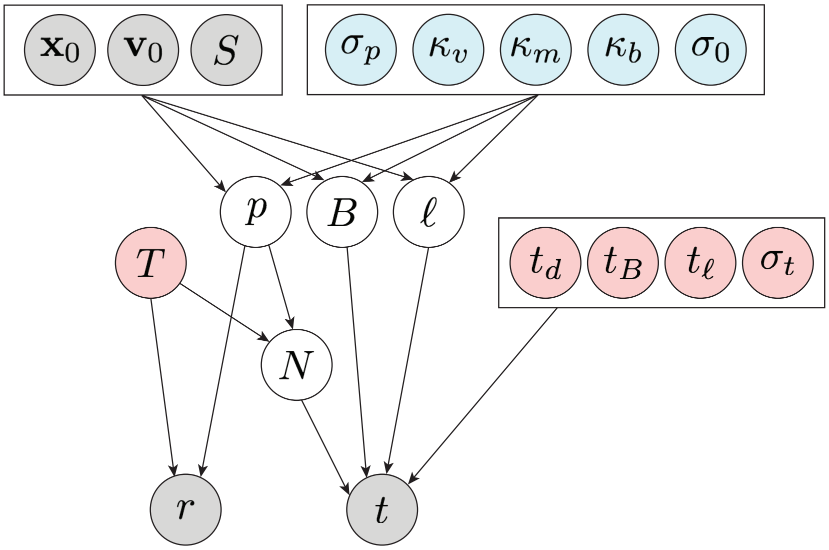


Figure 6.7: Full SPRT model. Gray circles are observed variables, blue circles are parameters fit to responses from a related experiment (see Appendix B.1), red circles are parameters fit to individual participants, and white circle are latent variables. Boxes are included for clarity only. The observed stimulus parameters ($\mathbf{x_0}$, $\mathbf{v_0}$, and $\mathbf{S}$) combine with the simulation parameters (σ_p, κ_ν, κ_m, κ_b, and σ_0) to determine the probability that the ball goes through the hole (p, Equation 6.3), the number of times it will bounce (B), and the distance it travels (ℓ). The probability that the ball goes in the whole, combined with the evidence accumulation threshold (T) determines the response (r, Equation 6.4) and the number of samples (N, Equation 6.5). The number of samples, bounces, and path length combine with the response time parameters (t_d, t_B, t_ℓ, and σ_t) to determine the response time (t, Equation 6.7).

which we approximate as:

$$p(t \mid T, p, \mu_B, \mu_\ell, t_d, t_B, t_\ell, \sigma_t) \approx \sum_{N=T}^{\infty} p(t \mid N, \mu_B, \mu_\ell, t_d, t_B, t_\ell, \sigma_t) p(N \mid T, p), \qquad (6.7)$$

where $\mu_B := \mathbb{E}_{\widetilde{\pi}}[B]$ and $\mu_\ell := \mathbb{E}_{\widetilde{\pi}}[\ell]$.

6.3.1.5 *Fitting the model*

To compare our model to the data collected in Experiments 6.1-6.3, we fit the parameters t_d, t_B, t_ℓ, σ_t, and T to the model defined by Equations 6.4 and 6.7 (and shown in Figure 6.7) using maximum *a posteriori* (MAP) estimation for each participant in each experiment. To compute the MAP estimate, we used standard Laplace priors for t_d, t_B, and t_ℓ; a Jeffrey's prior for the scale parameter of a Gaussian for σ_t; and a uniform categorical distribution over T. For each participant, we fit these parameters to only half the participants' responses, selected at random, and then computed all analyses (reported in the next section) on the remaining half of the data. As discussed in Appendix B.1, the parameters of the simulation model (σ_p, κ_v, κ_m, κ_b, and σ_0) were fit separately to the data from Experiment B.1. To compute p, μ_B, and μ_ℓ, we ran 10000 samples from the simulation model for each stimulus.

6.3.2 Results

6.3.2.1 Responses and response times

We first looked at how well our model was able to explain people's responses and response times overall. We computed Pearson correlations between the model and people with 95% confidence intervals computed from 10000 bootstrap samples. As shown in Figure 6.8, we found that the SPRT model achieved a high correlation with both people's responses and their response times across all three experiments. In Experiment 6.1, the model had a correlation of $r = 0.78,\ 95\%$ CI $[0.74, 0.82]$ with people's responses; in Experiment 6.2, this correlation was $r = 0.77,\ 95\%$ CI $[0.73, 0.81]$; and in Experiment 6.3, it was $r = 0.73,\ 95\%$ CI $[0.68, 0.77]$.

In Experiments 6.1 and 6.2, the model had roughly equal correlation with people's response times (Experiment 6.1: $r = 0.69,\ 95\%$ CI $[0.64, 0.73]$; Experiment 6.2: $r = 0.66,\ 95\%$ CI $[0.61, 0.72]$). In Experiment 6.3, the correlation with response times was lower, at $r = 0.48,\ 95\%$ CI $[0.40, 0.55]$. This is unsurprising, however: when analyzing the results of Experiment 6.3, we found that people were taking nearly a constant amount of time to respond for each stimulus. This implies that the majority of the variation in response times is attributable to noise and individual differences. Indeed, if we compute the correlation between the average z-scored response times of one half of the participants versus the those of the other half (chosen at random, with 1000 bootstrap samples), we find that the participants in Experiment 6.3 are barely predictive of each other: $r = 0.17,\ 95\%$ CI $[0.11, 0.25]$. The fact that the model has a higher correlation with participants than participants do with each other is not due to overfitting, because this correlation was computed using held-out data (which was not used for parameter fitting). Rather, the model results in a higher correlation because it is able to capture individual differences across participants.

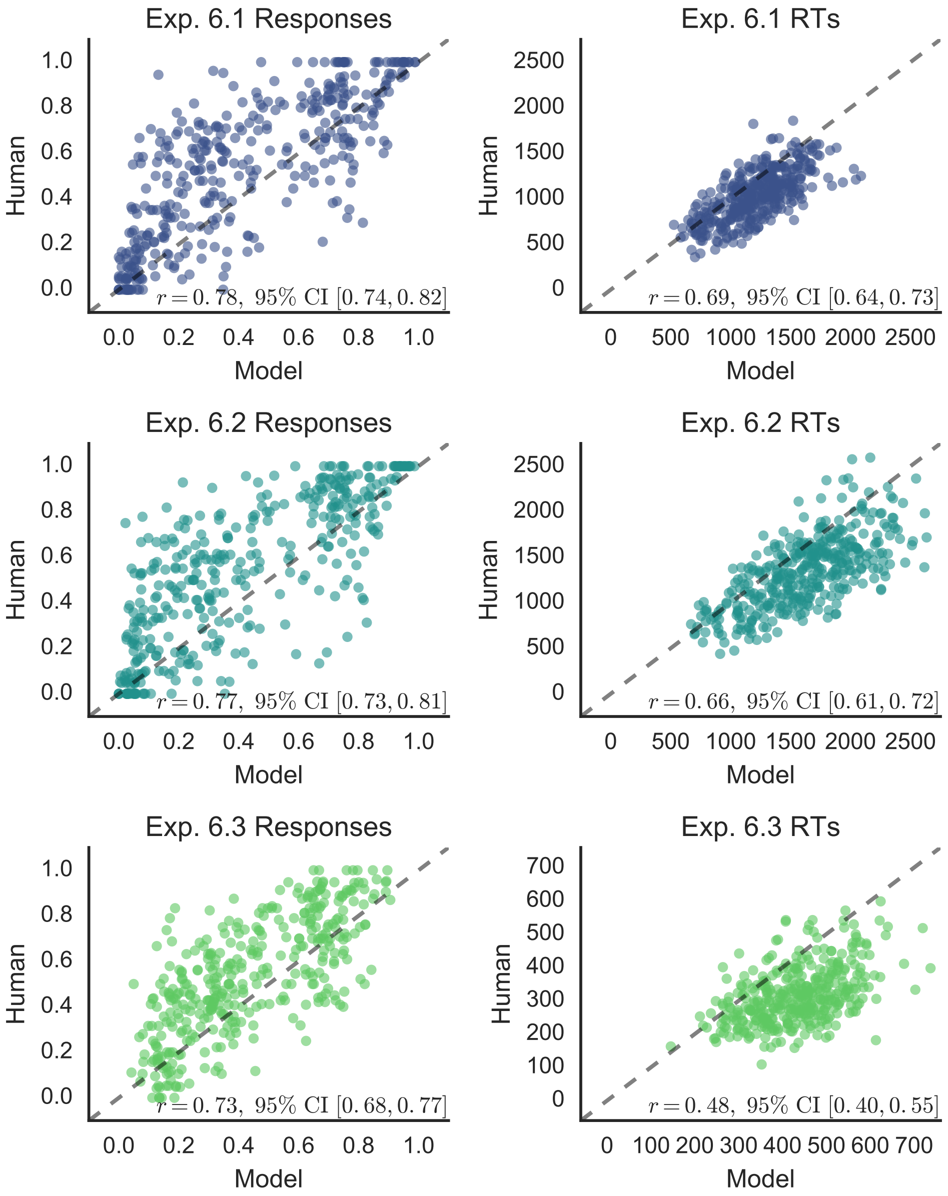


Figure 6.8: Model vs. human comparison. In all plots, each point corresponds to a different stimulus, trial type, and hole size. Dashed lines indicate perfect correspondence between model and people. Each row corresponds to data from a different experiment. *Left column:* The x-axis is the probability the model says the ball will go in the hole, and the y-axis is the proportion of participants saying the ball will go in the hole. *Right column:* Color and shape indicate the number of times the ball bounced during feedback. The x-axis is the model RTs, and the y-axis is the logarithmic mean RTs of participants, in milliseconds.

Table 6.1: Log posterior probabilities for different values of T. Bolded values indicate the lowest AIC or BIC, corresponding to the best fitting value of T.

		$T = 0$	$T = 1$	$T = 2$	$T = 3$
Exp. 6.1	AIC	323174	312966	**309118**	353557
	BIC	323925	314092	**310620**	355060
Exp. 6.2	AIC	370504	354595	**351117**	393930
	BIC	371258	355726	**352625**	395438
Exp. 6.3	AIC	**210370**	215324	244257	327789
	BIC	**211120**	216448	245755	329287

6.3.2.2 Evidence accumulation thresholds

We next examined which version of the SPRT model (as determined by the threshold parameter) people were most likely to use. As shown in Figure 6.9 (left), we found that $T = 2$ was the best fitting threshold for the plurality of participants. While there was no difference in the distribution of thresholds between Experiments 6.1 and 6.2 ($\chi^2(6) = 4.999, p = 0.54$), there was a difference in distributions between Experiments 6.1 and 6.3 ($\chi^2(3) = 53.085, p < 0.001$). In particular, while only 32.5% of participants in Experiment 6.1 had $T < 2$, the majority of participants in Experiment 6.3 (52.7%) had $T < 2$. Additionally, as shown in Table 6.1, a threshold of $T = 2$ overall fit participants' responses better in Experiments 6.1 and 6.2, while in Experiment 6.3, thresholds of either $T = 0$ or $T = 1$ fit participants' responses better than $T = 2$ according to measures of both AIC and BIC.

6.3.2.3 Fitted parameter values

We also looked at the average values of the fitted coefficients in order to further understand the differences in response times between the three experiments. The results are shown in Figure 6.9 (right). Interestingly, rather than finding a constant increase in parameter values between Experiments 6.1 and 6.2 (which had almost identical response characteristics, with the exception of an overall increase in response time), we found that there was a significant difference between the parameters for path length ($t(1) = -5.33, p < 0.001$) and response noise ($t(1) = -3.35, p < 0.01$), but not for non-decision time ($t(1) = 0.35, p = 1.00$) or the time to resolve a bounce ($t(1) = -0.01, p = 1.00$), with p-values adjusted for multiple comparisons using the Bonferroni method. Specifically, participants in Experiment 6.1 took substantially less

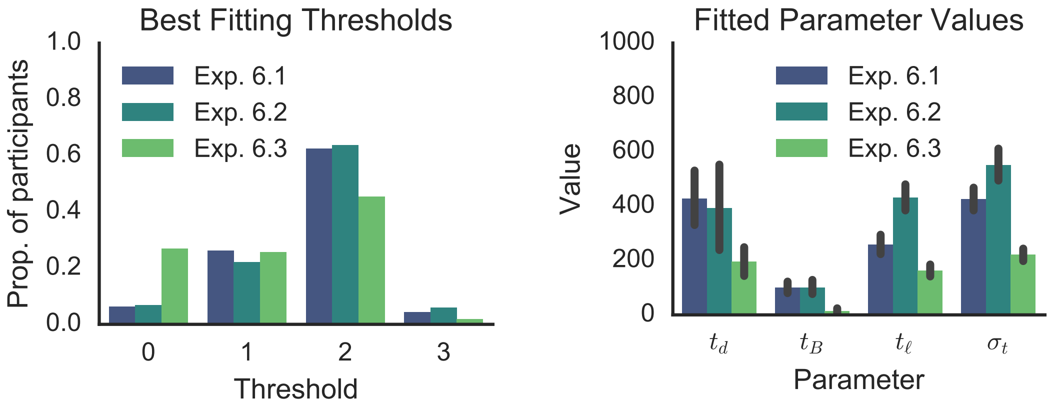


Figure 6.9: SPRT model fitted parameter values. *Left*: this subplot shows the proportion of participants best fit to particular threshold values, for thresholds of $T = 0, 1, 2,$ and 3. *Right*: this subplot shows the fitted values of the parameters in each experiment, averaged across participants, for non-decision time (t_d), the time to resolve a bounce (t_B), the time to run a single simulation (t_ℓ), and variance in response time (σ_t). Error bars indicate 95% confidence intervals. Note that t_B is multiplied by μ_B (approximately 0, 1 or 2) and t_ℓ is multiplied by μ_ℓ (approximately 1.25).

time to simulate the path of the ball ($t_\ell = 173.90,\ 95\%$ CI $[147.39, 201.90]$) than participants in Experiment 6.2 ($t_\ell = 306.78,\ 95\%$ CI $[267.67, 348.56]$). These results suggest that, rather than adjusting their threshold to take more samples, participants in Experiment 6.2 appear to have taken longer on each individual simulation. Participants in Experiment 6.3 appear to have taken less time across the board, with lower values of t_d, t_B, and t_ℓ.

6.3.3 Discussion

We constructed a model based on a combination of noisy physical simulation (Smith & Vul, 2013) and the sequential probability ratio test, or the SPRT (Wald, 1947). We found that our model fit participants responses and response times from Experiments 6.1-6.3 extremely well, and revealed that people's behavior corresponds to a sampling strategy of mental simulation with a very low evidence accumulation threshold. Specifically, we found in Experiments 6.1 and 6.2 the best fitting threshold to be $T = 2$, corresponding to an average ranging between 2 samples (when $p = 0$ or $p = 1$) to 4 samples (when $p = 0.5$). In Experiment 6.3, we found the best fitting threshold to be $T = 1$, corresponding to taking a single sample every time, revealing that the increased time pressure caused people to lower their threshold.

6.4 General Discussion

In this chapter, we asked whether people adapt the amount of computation they perform with mental simulation depending on how difficult the task is. Across three experiments, we found broad empirical support for the hypothesis that they do, demonstrating that people take longer to respond when they are more uncertain. We additionally found that the amount of time people take to respond is mediated by the payoff structure of their environment; however, the increased response time may not necessarily result in better performance on the task. We modeled people's behavior using a model which combined noisy physical simulation (Smith & Vul, 2013) with a decision-making strategy known as the sequential probability ratio test, or the SPRT (Wald, 1947). This model captured both responses and response times, revealed that people run very few mental simulations before making a judgment; showed that people can change the amount of adaptivity in their responses (e.g., by lowering the evidence accumulation threshold down to $T = 1$ when they are under time pressure); and indicated that people may be able to adjust the amount of time they spend on each individual simulation (as evidence by differences in values for t_ℓ across experiments).

One criticism of the hypothesis that people use mental simulation to reason about physical scenarios is that simulation is far too computationally intensive to make rapid physical inferences (Davis & Marcus, 2014); indeed, models of mental simulation often use hundreds or thousands of samples to compute their predictions. While some post-hoc analyses have suggested that participants in other experiments may only take a handful of samples (e.g., Chapter 4; Battaglia et al. (2013)), there has been no previous study aimed at determining exactly how many mental simulations people run. Here, we present strong evidence that people rely on a small number of simulations, corroborating theoretical predictions that taking a small number of samples is actually optimal behavior (Vul et al., 2014). Importantly, these results suggest that mental simulation may not be so intractable after all.

In the remainder of this chapter, we discuss alternate hypotheses and models that might be able to account for the results presented here.

6.4.1 Alternate Hypotheses

There are a number of alternate hypotheses that might be able to explain people's behavior which we have so far not yet considered. For example, a key component of our model is that the simulations are *sequential*. However, could it be the case that people are running multiple simulations in parallel? We argue that the answer is no: multiple simulations in parallel would lead to large differences in accuracy, but not response time. Our results show just the opposite: people exhibit significant differences in response time.

A more competitive alternate hypothesis is that people are not running multiple simulations at all, but that they run a single simulation with a variable level of granularity. Under this hypothesis, the idea would be that people form an estimate of the difficulty of the task, and then run detailed simulations for difficult scenes versus coarse simulations for easy scenes. Some of our empirical results suggest this could potentially be the case: for example, in Experiment 6.2, people increased the time they spent simulating each sample compared to Experiment 6.1. This increase in simulation time could perhaps be attributable to running a more detailed simulation. Unfortunately, with only the empirical data reported in this chapter, it would difficult to distinguish between these two hypotheses; a more rigorous test would require additional behavior measures such as eye-tracking data.[1]

There are a few reasons why we might think the multiple sample hypothesis is *a priori* more plausible than the granularity hypothesis. First, the SPRT model provides an optimal account for why response times have long tails, in that they result from occasionally taking many samples. If the granularity were chosen according to a long-tailed distribution (e.g. gamma or chi-squared), for example, and we assume that response time is inversely proportional to granularity, then we would expect response times to also have a long tail (distributed according to, e.g. inverse gamma or inverse chi-squared). Yet, why would granularity be chosen according to a long-tailed distribution in the first place? Second, if the granularity is related to the difficulty of the problem, how would people know which granularity to use without running simulations? It is possible that people might be able to make a guess about the difficulty based on visual heuristics (for example, whether the ball is approximately headed towards the hole or not), but it does not seem likely that such guesses would very accurately track the true difficulty of the problem.

In our view, the most likely hypothesis is that people are *both* adapting the number of samples and varying the granularity of their simulations. The differences in values for t_ℓ between the three experiments provide some preliminary evidence for this hypothesis. Additionally, if we take a closer look at the distributions of response times predicted by the SPRT model versus the empirical distributions, we find they do not perfectly match. Figure 6.10 shows quantile-quantile (QQ) plots for the log response times generated by the model, versus the log response times

[1] Eye-tracking data from a similar experiment by Gerstenberg, Peterson, et al. (2017) does reveal that people make multiple saccades to a future mentally simulated path of a ball. However, further analysis would be needed to determine whether the saccades correspond to multiple simulations, or whether they are part of the same simulation.

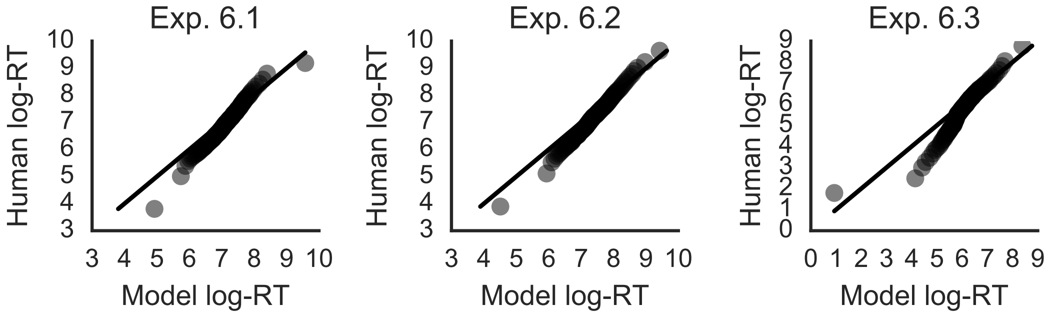


Figure 6.10: Quantile-quantile plots of log response times. Each subplot shows a QQ plot of model log response times versus human log response times for each experiment. The solid line indicates perfect correspondence.

generated by people. These plots reveal that human RTs are skewed more strongly towards low values than the model's response times, and that they also have slightly heavier tails at the other end of the distribution. This mismatch is perhaps reflective of additional degrees of freedom that we have not captured in our model: for example, that people might be changing the granularity of their simulations on a trial-by-trial basis. This difference might also reflect trial-by-trial variation in the evidence accumulation threshold, in contrast to having a fixed threshold for the entire experiment as we have modeled it here. Future research will be needed to tease these various possibilities apart.

6.4.2 Comparison of the sprt and the Drift-Diffusion Model

One might object to our use of the sequential probability ratio test as the basis for our model, rather than more well-known models of decision making such as the drift-diffusion model, or ddm (Ratcliff & McKoon, 2008). However, as mentioned in the introduction, the sprt is actually very closely related to the ddm; in fact, it is the discrete analogue to the ddm. As discussed by Bogacz et al., "as discrete samples are taken more frequently and one approaches sampling of a continuous variable, the sprt converges on the ddm" (pg. 703-4, Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006).

Because of their mathematical relationship, we do not view the sprt and the ddm as competing models, but rather as two variations of the same model. Depending on the decision-making process being modeled, it may be more appropriate to use the sprt or the ddm. For perceptual decision making tasks where the accumulation of perceptual evidence is rapid and continuous, it makes sense to use the DDM. For cognitive decision making tasks where there is no continuous source of evidence—but where it may be possible to take discrete samples, such as via mental simulation—it makes sense to use the SPRT.

6.4.3 Conclusion

Mental simulation is a powerful and flexible tool, as it offers a way to make predictions about scenarios that have not yet (or may never) come to pass. In this work, we demonstrated that when people use mental simulation, they are sensitive to their own uncertainty in reasoning about the task and accordingly adjust how many simulations they run. These results are among the first to explain not just *that* people use simulation to reason about the world, but *how* they use it. While there are still many questions left unanswered—e.g., how do people use simulations in non-binary tasks?—this work brings us one step closer to understanding of how mental simulation is used.

Without leaps of imagination or dreaming, we lose the excitement of possibilities. Dreaming, after all, is a form of planning.

Gloria Steinem

7

A Formal Framework for Modeling Mental Simulation

When trying to understand how people solve a particular problem, it is tempting to think that "mental simulation" is an answer in and of itself. Such a singular answer is so seductive because it feels so natural: when you run a mental simulation, you somehow just *know* which simulation(s) to run, how much time to spend simulating, and what conclusions to draw. You are likely not even aware of making these decisions; it is the phenomenological aspects of the simulation (i.e., mental imagery) that draw your attention. However, as I have hopefully convinced you through the last few chapters, such decisions are important components of the mental simulation process and cannot be ignored if we desire to have a satisfying cognitive theory of how people solve problems through mental simulation.

In this chapter, I argue that the meta-level questions surrounding mental simulation—such as which simulations to run or how long to run them for—have been neglected because we have not had the framework through which to view mental simulation. First, I show how the language of partially observable Markov decision problems (Kaelbling, Littman, & Cassandra, 1998), or POMDPs, provides a template for thinking about the problems that mental simulations solve at the computational level of analysis (Marr, 1982). This framework captures the structure of problems that have been shown to be solved by mental simulation in the literature, and also generalizes to new types of mental simulation problems that have not been investigated in as much detail. Second, I argue that POMDPs in and of themselves do not sufficiently capture the meta-level decisions made about mental simulation. To understand such decisions, we necessarily must focus attention at the algorithmic level of analysis; I suggest that this may be accomplished through a resource-rational analysis (Griffiths et al., 2015) on top of the POMDP framework. The resource-rational analysis allows us to understand *why* mental simulation is so pervasive, as well as providing hypotheses for how the mind is able to use it so flexibly.

7.1 Computational-Level Analysis

As discussed in Chapter 3, the process of mental simulation can be defined as either a transition function or an observation function. Here, I use this definition to define not just what mental simulation itself is, but what are the *problems* it is trying solve.

In its most general form, a problem that may be solved through mental simulation can be defined at the computational level of analysis (Marr, 1982) using the terminology of a *partially observable Markov decision problem*, or POMDP (Kaelbling et al., 1998). A POMDP (Figure 7.1a) is described as a tuple $\langle\mathcal{S}, \mathcal{A}, T, R, \Omega, O\rangle$, where $s \in \mathcal{S}$ are states, $a \in \mathcal{A}$ are actions, $T : \mathcal{S}\times\mathcal{A} \rightarrow \Pi(\mathcal{S})$ is the transition function mapping states and actions to a probability distribution over new states, $r \in \mathbb{R}$ are rewards, $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function mapping states and actions to rewards, $o \in \Omega$ are observations, and $O : \mathcal{S} \rightarrow \Pi(\Omega)$ is the observation function mapping states to a probability distribution over observations.

Typically, the "problem" component of a POMDP is to maximize the expected discounted future reward, $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where r_t is the reward received at time t. For our purposes, we need not always focus on the reward maximization aspect.[1] For example, for some problems, the goal might be to infer the underlying states based on the observations. In others, it might be to infer the actions which gave rise to the observed set of states. In this way, by focusing on different aspects of the POMDP, it is easy to formulate the computational-level problem that mental simulation solves across many different tasks. In the following subsections, I describe a

[1] Technically, any learning problem with discrete time steps can be viewed as a RL problem where the reward to be maximized is a reward for choosing the right answer. However, when the learning problem does not involve a sequence of actions and rewards, it is often easier to solve the problem using standard inference or supervised learning techniques. An interesting middle ground arises when the problem is technically an inference problem, but where actions may be taken to gather information about the solution: for example, by inducing collisions between objects to gauge their relative masses (Bramley, Gerstenberg, & Tenenbaum, 2016). RL has been proposed as a way of choosing such actions in the artificial intelligence community (Denil et al., 2017), though this has not yet been evaluated as a model of human cognition.

few different types of such problems, and show the subset of the POMDP which is relevant to the problem in Figure 7.1.

7.1.1 State inference

The mass inference task from Chapter 4 is an instance of a state inference task. Here, the rewards $\mathcal{R}$ and actions $\mathcal{A}$ are ignored, and the focus is on inferring a property of the states given observations. Specifically, if one assumes states are tuples $s_t := \langle x_t, \rho \rangle$, where x is the observable position and orientation of the blocks in the tower, and ρ is the unobserved physical property of mass, then the computational-level problem is to infer:

$$p(\rho \mid \mathbf{o}_{1:T}) \propto \int_{\mathbf{x}_{1:T}} \prod_{t=1}^{T} p(o_t \mid x_t, \rho) p(x_t \mid x_{t-1}, \rho) p(\rho) \, \mathrm{d}\mathbf{x}_{1:T}. \tag{7.1}$$

A graphical depiction of this inference problem is shown in Figure 7.1b.

7.1.2 Path planning

The mental rotation task from Chapter 5 is an instance of a path planning task. An agent is given observations of the shapes, and must determine what transformation is required to bring the shapes into alignment (either a rotation, or a rotation and a reflection). These transformations correspond to the actions of the POMDP; however, there are no rewards for taking any particular action (though there is a reward for achieving the goal).[2] Then, the computational-level problem is to infer the actions that link the two observations:

$$p(\mathbf{a}_{1:T-1} \mid o_1, o_T) \propto \int_{\mathbf{s}_{1:T}} p(o_1 \mid s_1) p(o_T \mid s_T) \prod_{t=1}^{T-1} p(s_{t+1} \mid s_t, a_t) p(a_t \mid s_t) \, \mathrm{d}\mathbf{s}_{1:T}, \tag{7.2}$$

where T is not a fixed value. Given the inferred actions, the answer to "same" or "flipped" can be determined by computing a function over inferred actions (e.g., an odd number of "reflect" actions means they are flipped, and an even number means the are the same). A graphical depiction of this inference problem is shown in Figure 7.1c.

[2] Often in tasks like path planning, it is assumed that there is a small cost for each action. However, in the computational-level formulation of this problem, the cost is irrelevant: it is only important when considering resource constraints such as time or metabolic cost. I come back to this point when discussing the resource-rational analysis of the path planning problem.

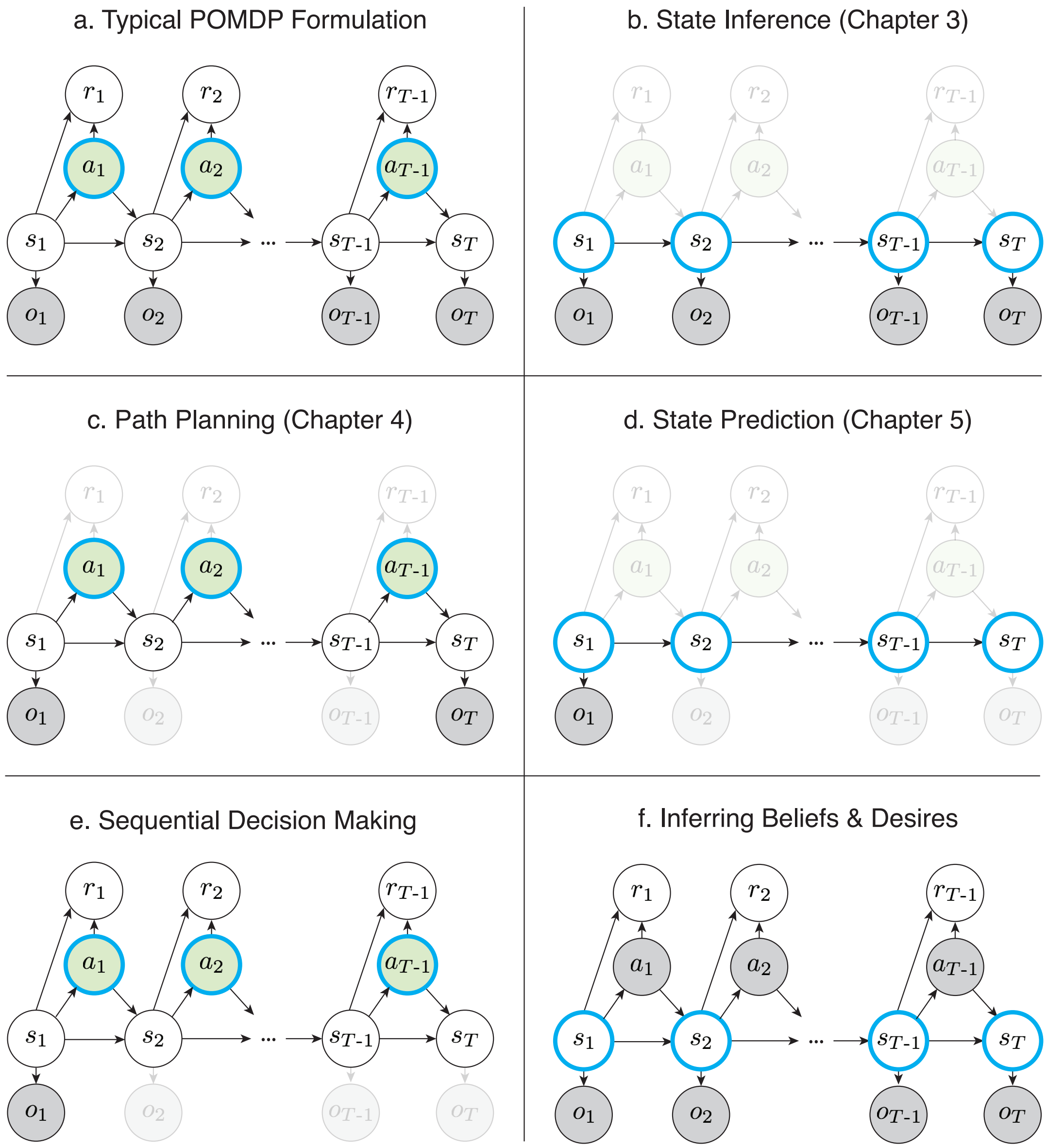

a. Typical POMDP Formulation
b. State Inference (Chapter 3)
c. Path Planning (Chapter 4)
d. State Prediction (Chapter 5)
e. Sequential Decision Making
f. Inferring Beliefs & Desires
r_1
r_2
$r_{T\text{-}1}$
a_1
a_2
$a_{T\text{-}1}$
s_1
s_2
...
$s_{T\text{-}1}$
s_T
o_1
o_2
$o_{T\text{-}1}$
o_T

Figure 7.1: POMDP formulation of mental simulation tasks. The gray circles indicate observed variables. The green circle indicate variables that can be intervened on (actions). Dimmed areas indicate variables that are not relevant to a particular task. Bold blue outlines indicate variables that are inferred for a particular task. The variables correspond to rewards (r), actions (a), states (s), and observations (o).

7.1.3 State prediction

Predicting whether a ball will go through a hole as in Chapter 6 is an instance of a state prediction task. There is only a single observation of the initial state of the ball, from which the final state of the ball must be predicted:

$$p(\mathbf{s}_{1:T} \mid o_1) \propto p(o_1 \mid s_1) \prod_{t=1}^{T} p(s_t \mid s_{t-1}). \tag{7.3}$$

The answer to whether the ball goes through the hole can be computed based on the sequence of states (e.g., whether there exists one state where the ball is one side of the wall with the hole in it, and another state whether the ball is on the other side of the wall). A graphical depiction of this inference problem is shown in Figure 7.1d.

7.1.4 Sequential decision making

There are even more sophisticated mental simulation tasks than those presented in this thesis. For example, in sequential decision making tasks such as game playing (van Opheusden, Bnaya, Galbiati, & Ma, 2016), mental simulations are used to plan actions, as shown in Figure 7.1e. This formulation is much closer to the original POMDP, and is identical to a MDP with the exception that the initial state may not be directly observed. The goal is to choose actions which will maximize the expected sum of future rewards. The value of taking action a_t while in state s_t is given by the Bellman equation (Bellman, 1957):

$$Q(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{s_{t+1}} p(s_{t+1} \mid s_t, a_t) \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}), \tag{7.4}$$

where $0 \leq \gamma \leq 1$ is a discount factor. Given the optimal Q-function, the optimal action to take in state s_t is given as $a_t = \arg\max_a Q(s_t, a)$. As the initial state is unobserved, choosing a_1 is slightly different:

$$a_1 = \arg\max_a \sum_{s_1} p(s_1 \mid o_1) Q(s_1, a).$$

The computational-level problem is then to compute Q, from which all a_t can be directly determined.

7.1.5 Inferring beliefs and desires

Mental simulation may also be used to reason about the behavior of other agents, and in particular, to infer their goals and values based on observed actions (Figure 7.1f). This formulation also naturally falls within our framework, and is typically referred to in the cognitive science literature as *Bayesian theory of mind*, or BToM (Baker & Tenenbaum, 2014). In what follows, I briefly reproduce the BToM formulation as given by Baker and Tenenbaum (2014).

In BToM, the states s are factored into separate components: s^A, representing the state of the observed agent, which is typically observed; s^W, representing the state of the environment or world, which may or may not be observed;[3] and s^D, representing the observed agent's desires, which is not observed by the observer (but which is known to the observed agent). The agent's desires affect the reward function,[4] $R(s, a)$. It is assumed that the agent maintains a belief over s^W, denoted as $b \in \Pi(\mathcal{S}^W)$, and acts according to the Q-function:

$$\begin{aligned}Q(s_t^A, s^D, b_t, a_t) = \sum_{s^W} b_t \cdot \Big[R(s_t^A, s^D, a_t) + \gamma \sum_{s_{t+1}^A} p(s_{t+1}^A \mid s_t^A, a_t) \cdot \\ \sum_{o_{t+1}} p(o_{t+1} \mid s_{t+1}^A, s^W) \max_{a_{t+1}} Q(s_{t+1}^A, s^D, b_{t+1}, a_{t+1})\Big],\end{aligned}$$

where b_{t+1} is computed according to Bayes' rule:

$$b_{t+1} \propto p(o_{t+1} \mid s_{t+1}^A, s^W) p(s_{t+1}^A \mid s_t^A, a_t) \cdot b_t. \tag{7.5}$$

The computational-level problem is to infer the agent's beliefs $\mathbf{b}_{1:T}$ and their desires s^D based on their observed behavior. Assuming the agent takes actions according to a distribution which is a function of the Q-function (such as a softmax), we have:

$$\pi(a_t \, | s_t^A, s^D, b_t) = f(Q(s_t^A, s^D, b_t, a_t)),$$

[3] Parts of the world are often unobserved; for example, Baker and Tenenbaum (2014) use the example of not knowing the location of a set of food trucks on any particular day.

[4] This reward function can be further decomposed into multiple terms reflecting different aspects of the agents: for example, that agents both receive a benefit for achieving goals, as well as incur a cost for taking different types of actions (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016).

Given this distribution, the beliefs may be inferred as:

$$p(\mathbf{b}_{1:T}, s^D \mid \mathbf{a}_{1:T}, \mathbf{s}^A_{1:T}, s^W, \mathbf{o}_{1:T}) \propto \\ p(s^D)\prod_{t=1}^{T} p(o_t \mid s^A_t, s^W)p(s^A_t \mid s^A_{t-1}, a_{t-1})\pi(a_{t-1} \mid s^A_{t-1}, s^D, b_{t-1})\delta(b_t), \tag{7.6}$$

where $\delta(b_t) := \delta(b_t \mid o_t, s^A_t, s^A_{t-1}, a_{t-1}, b_{t-1})$ is 1 when b_t is equivalent to that computed by Equation 7.5 and 0 otherwise.

7.1.6 Other learning problems

The POMDP framework allows us to also consider other classes of learning problems which have largely not been considered when modeling mental simulation. To do this, one can focus on different subsets of the graph in Figure 7.1a. For example, consider a combination of the state inference and path planning problems (Figure 7.1b-c) where $\mathbf{o}_{1:T}$ are observed, $\mathbf{s}_{1:T}$ are unobserved, and $\mathbf{a}_{1:T}$ need to be inferred. In more concrete terms, the agent sees some observations that are a consequence of actions being taken, and must infer what those actions are. This type of problem aligns naturally with the types of phenomena observed in the "mirror neuron" literature, in which the brain appears to make inferences about the underlying actions being taken by another agent based only on visual observation (e.g. Rizzolatti, Fadiga, Gallese, & Fogassi, 1996).

7.2 Resource-Rational Analysis

In the previous section, I outlined how a number of psychological behaviors can be interpreted at the computational level of analysis (Marr, 1982) using the language and framework of POMDPs (Kaelbling et al., 1998). The behaviors described often involve some amount of mental simulation, which can be viewed as the unrolling of the the transition function, $T(s, a)$. However, many of the computational-level problems described in the previous section *need* not be solved in this way. For example, in the case of state inference, if $\mathbf{x}_{1:T}$, $\mathbf{o}_{1:T}$, and ρ are all jointly Gaussian, then Equation 7.1 can be computed analytically and there is no need to run any simulations. Or, in the case of sequential decision making, the Q-function may be learned using model-free reinforcement learning algorithms such as TD learning (Sutton, 1988). Such model-free RL algorithms do not require the use of simulations, and in fact assume that the transition function $T(s, a)$ is unknown.

The observation that mental simulation is not *required* to solve problems such as those listed in the previous section is not new (Davis & Marcus, 2014), but it does raise a number of difficult questions. First, what is it about simulation that makes it such a frequently used cognitive ability? Second, how can we (as psychologists) hope to predict when someone might use mental simulation as opposed to an alternate strategy (such as exact inference, or a heuristic) to solve the computational-level problems listed above? And third, how can we characterize the meta-level decisions that people make when using mental simulation? I argue that all of these questions may be answered by considering the fact that people have limited computational resources, through the approach of *resource-rational analysis* (Griffiths et al., 2015).

7.2.1 An overview of resource-rational analysis

The key idea behind resource-rational analysis is to analyze human cognition first at the computational level, and second at the algorithmic level through approximate solutions to the computational-level problem (Griffiths et al., 2015). Many computational-level problems are intractable and can only be solved approximately; however, there are often many different methods for approximation. For example, Bayesian inference can be approximated via simple Monte Carlo, Metropolis-Hastings (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), Gibbs sampling (Geman & Geman, 1984), slice sampling (Neal, 2003), Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, & Roweth, 1987; Neal, 2011), and variational inference (Blei, Kucukelbir, & McAuliffe, 2017; Peterson & Anderson, 1987), just to name a few options. Different problems are well-suited to different inference methods, and the choice of method is usually determined by the modeler whose skill and experience inform the decision. To complicate things further, each method has its own hyperparameters; for example, in Metropolis-Hastings, choices need to be made about the proposal distribution, the number of burn-in samples, the thinning rate, the number of chains, and the length of chains. As with the choice of the approximation itself, the choice of hyperparameters are usually determined by the modeler and their efficacy by the modeler's expertise.

Resource-rational analysis suggests that agents choose approximate solutions to computational-level problems—and correspondingly set their hyperparameters—by optimizing with respect to the agents' resource limitations. These resource limitations may range from more computational-level constraints (i.e., not having access to all the data) to algorithmic- and implementation-level concerns (i.e., neural firing rates or working memory capacity). A particular resource-rational analysis needs to decide which of these resource constraints are relevant for the particular situation. This pushes the decisions required by the experimenter to a more abstract level (i.e., determining what variables are relevant to the problem) rather than the level of architectural choices of the algorithm itself (i.e., setting hyperparameter values).

To summarize the approach suggested by resource-rational analysis, I paraphrase the steps given by Griffiths et al. (2015, p. 223) here. The first step is the same as in regular rational analysis (Anderson, 1990): determine the computational-level problem and the optimal solution (e.g., Bayesian inference). The second step is to consider the class of algorithms which gives approximations to the optimal solution of the computational-level problem (e.g., Metropolis-Hastings). Finally, the third step is to solve the trade-off between resource usage and performance.[5] Here, "resource usage" may be any resources the mind might care about: time, metabolic cost, etc. "Performance" is how well the mind is able to actually solve the problem; for example, approximation error might serve as a proxy for performance.

In the sections that follow, I describe how this approach can help to answer the three questions posited at the beginning of this section. While I do not always follow the three steps described above to the letter, I do maintain the spirit of resource-rational analysis, which is to take seriously the idea that the agent has limited computational resources.

7.2.2 Why mental simulation?

First, what is it about simulation that makes it such a frequently-used cognitive ability? The common refrain is that having a flexible and general strategy such as mental simulation allows us to reason about a wide range of scenarios, even novel ones that we have not previously encountered. Simpler strategies tend to be brittle in that they only apply to a small subset of possible situations. Often, this criticism is leveled at heuristics (Gigerenzer & Brighton, 2009), but it may be true even of optimal strategies (e.g. Belousov et al., 2016). This "generalization" argument has been made many times in the cognitive science literature (e.g. Craik, 1943; Helmholtz, 1925; Tenenbaum et al., 2011), and is also recognized among machine learning researcers when discussing generative versus discriminative models (e.g. Ng & Jordan, 2002) or model-based versus model-free methods for reinforcement learning (e.g. Atkeson & Santamaria, 1997).[6]

There is an additional argument to be made as to why a general strategy like mental simulation

[5] Griffiths et al. (2015) also include a fourth step, which is to compare the derived model to human data and iterate on the first three steps as necessary.

[6] In machine learning, the "generalization" argument is usually presented in terms of "data efficiency": a generative method is more data efficient than a discriminative method, and model-free methods are data-inefficient compared to model-based methods. However, data efficiency can be thought of as a proxy for generalization: methods that have high data efficiency can generalize to new cases based on fewer data points than methods with low data efficiency.

might be necessary. In addition to mental simulation just being *useful*, it is also more *efficient* to have a small set of strategies which are together general enough to handle the wide range of situations that an agent might find themselves in. Specifically, if we consider that people are bounded in their computational resources, then it makes sense to only have a few strategies because an increasing number of strategies also increases the cost of metareasoning (Milli, Lieder, & Griffiths, 2017). If people had access to many possible strategies—say, hundreds or even thousands—then it would be likely that for any given scenario one of those strategies would be sufficient, even if all were relatively limited. However, if people only have access to a small number of strategies, then it becomes necessary for at least one of those strategies to be highly flexible and generalizable because there are simply not enough strategies to cover the full space of possibilities.

7.2.3 Strategy selection

Of course, mental simulation is often not the *only* strategy at hand; there may be both cheaper options (such as heuristics) and more expensive options (such as logically reasoning through a set of equations, in the case of physical reasoning) available. Indeed, there is ample evidence that people do not always rely on mental simulation (Schwartz & Black, 1996, 1999; Smith, Battaglia, & Vul, 2013). This brings us to the second question posed above: how can we (as psychologists) hope to predict when someone might use mental simulation as opposed to an alternate strategy? To answer this question, the computational-level problem can be formulated at the meta-level as the problem of determining which computation(s) to perform in order to satisfy some objective (such as a speed/accuracy trade-off). This is the same idea behind the general notion of metareasoning (Hay, Russell, Tolpin, & Shimony, 2012; Russell & Wefald, 1991) and has been used to explain how people choose between different strategies to solve a problem, such as which card-sorting algorithm to use or which heuristic to rely on (Lieder & Griffiths, 2017; Lieder et al., 2014).

To be more precise, let us say that there is a set of n functions $f_1, f_2, \ldots, f_n$ such that each function $f_i \in \mathcal{F}$ maps from inputs, $x \in \mathcal{X}$, to outputs, $o \in \mathcal{O}$: $f_i : \mathcal{X} \rightarrow \mathcal{O}$. The objective, which is referred to as the *value of computation*, is based on the utility of the output of the function as well as the resources taken to compute it. The utility, denoted as $u \in \mathcal{U}$, is given via a function, $U : \mathcal{F} \times \mathcal{X} \rightarrow \mathcal{U}$. The resources, denoted as $c \in \mathcal{C}$, are similarly given via a function, $C : \mathcal{F} \times \mathcal{X} \rightarrow \mathcal{C}$. The total VOC is:

$$\mathrm{VOC}(f_i, x) = \mathbb{E}[U(f_i, x) - \gamma \cdot C(f_i, x)], \tag{7.7}$$

where γ is a scaling factor. The goal, then, is to find the f_i which maximizes the VOC.[7] To give a more concrete example, let us say that the task is to sort a set of playing cards. Then, the different functions f_i would correspond to different sorting algorithms (e.g., cocktail sort, insertion sort, etc.) and the input x would correspond to a particular set of cards to be sorted. Such algorithms have different run times depending on what x is, and people may be more or less likely to make mistakes depending on the algorithm as well (Lieder et al., 2014).

The challenge in maximizing the VOC is that it is known neither how useful nor how expensive a given computation will be before actually executing it. In the worst case scenario, where there is no correlation between the input, x, and the utility or cost, it is impossible do better than choosing a strategy at random (or choosing the best overall strategy). Luckily for us, the world tends to be correlated and piecewise smooth. This means that two inputs, x and x', are likely to have similar utilities and costs if the inputs are in some sense "similar" to each other. Indeed, Lieder and Griffiths (2017); Lieder et al. (2014) showed that this assumption often holds by learning approximations to the functions $U(f_i, x)$ and $C(f_i, x)$ based on features $\phi^{(x)}$ of the input, such that $U(f_i, x) \approx \widehat{U}_i(\phi_1^{(x)}, \ldots, \phi_k^{(x)})$ and $C(f_i, x) \approx \widehat{C}_i(\phi_1^{(x)}, \ldots, \phi_k^{(x)})$. To return to our card sorting example (Lieder et al., 2014), the features ϕ would correspond to details of the input (e.g., the number of cards, a measure of how close to sorted they are, etc.).

The strategy selection framework described above can be naturally extended to mental simulation as well. To do so, consider mental simulation to be a particular f_i, and an alternate strategy (such as a heuristic) to be a different f_j. The difficult question is: how do should the features of the input space be defined? Such features may solely be visual properties of the scene under consideration. For example, Smith, Dechter, et al. (2013) discuss scenes in which participants have to predict whether a ball will reach a red goal or a green goal first; for many trials, people seem to use simulation, while on other trials (where the topology might prevent the ball from ever reaching one of the goals), people seem to use an alternate strategy.[8] In other cases, the features might be linguistic, visual, or motor features of the way the problem is presented. For example, it is a common finding that when people are asked to make physical predictions, they make systematic errors when the problem is presented verbally or schematically, but tend to make correct predictions when they are asked to take actions (Jennings & Davies, 2017; Schwartz & Black, 1999; Smith, Battaglia, & Vul, 2013; Zago & Lacquaniti, 2005). This dichotomous behavior might be

[7] In the most general form of strategy selection, there would be a sequence of multiple actions, placing us in the realm of sequential decision making. This is the more general notion of metareasoning, described in further detail in Russell and Wefald (1991) and Hay et al. (2012).

[8] Though see Smith et al. (2017), who argue based on reaction times that simulation is still being used in these "containment" trials.

explained by a differences in the features used by metareasoning; it is an open area of research to explore whether this is in fact the case.

7.2.4 The "hyperparameters" of mental simulation

Finally, how can psychologists characterize the meta-level decisions that people make when using mental simulation? Just as with using metareasoning to solve the trade-off between utility and resource usage for different strategies, we can also use metareasoning to solve the same trade-off for different choices of hyperparameters. After all, there is not much fundamentally different between a space of strategies and a space of hyperparameters (beyond the fact that hyperparameters may lie in a multidimensional and continuous space, while the space of strategies is typically unidimensional and discrete). To use metareasoning to choose the hyperparameters, however, we first must specify what the hyperparameter spaces even *are*. To do this, I would first like to distinguish between two types of hyperparameters: *representational* hyperparameters, and *meta* hyperparameters.

Representational hyperparameters affect the representation of the simulation itself. For example, these parameters might include the granularity of the simulation (i.e., the timestep), the geometric representation of the objects (in the case of a physical simulation), and the uncertainty or noise in the simulation. The choice of representational hyperparameters is quite difficult in that it is often not obvious whether the chosen representation of the simulation is correct or not. For example, while in Chapter 4 we assumed a physical simulation of 10 objects with perceptual and force uncertainty, this representation may only be a coarse approximation of the representation that people are actually using. It is likely, for example, that they do not run a simulation of all 10 objects at once, particularly given the difficulty that people have in tracking more than a few objects simultaneously (Pylyshyn & Storm, 1988). Representational hyperparameters may play a large role in metareasoning if they affect the quality or length of the simulation; however, they are very difficult to pin down exactly.

Regardless of what the representation of the simulation is, it is still necessary to make choices about meta-hyperparameters such as how many simulations to run and how long to run them for. These meta-hyperparameters also affect metareasoning, but are perhaps somewhat easier to analyze in that they are largely independent of the specific representation. This is not to say that meta-hyperparameters cannot interact with representational hyperparameters: they can. For example, if a coarse representation is chosen, then more simulations will need to be run to achieve the same level of accuracy. However, assuming the representational hyperparameters remain fixed, we can reason about what the meta-hyperparameters are; this is exactly the approach taken in Chapter 6.

To be more precise, we can rewrite Equation 7.7 to reflect the choice of hyperparameters instead of the choice of algorithm:

$$\mathrm{VOC}(f, x, \theta) = \mathbb{E}[U(f, x, \theta) - \gamma \cdot C(f, x, \theta)], \tag{7.8}$$

where θ are the hyperparameters of the simulation, f. The goal now is to find the θ which maximize Equation 7.8. This can either be solved by approximating the VOC, as in the previous section, or turn to specific algorithms which maximize the VOC in other ways. For example, the SPRT (Wald, 1947) solves Equation 7.8 by assuming that the parameters correspond to the number of samples, that the utility corresponds to accuracy, and that the cost corresponds to time. The SPRT then yields the minimum cost solution for a given level of accuracy (Wald & Wolfowitz, 1948).

7.2.5 Examples

In what remains of this chapter, I go through several of the example problem types depicted in Figure 7.1 and discuss how resource-rational analysis can be applied to the POMDP formulation of the problem. Through these case studies, I demonstrate how this approach reveals similarities between how mental simulation is used across different scenarios, and how it can generate new hypotheses and modeling approaches.

7.2.5.1 State prediction and inference

Although Equation 7.1 appears complicated, it can be simplified if we assume that the observations are the feedback F (i.e., whether the tower falls) which is a function of the entire simulation $\mathbf{x}_{1:T}$, and the initial observed configuration c_0. This allows us to rewrite the equation as:

$$p(\rho \mid F, c_0) \propto p(\rho) \int_{\mathbf{x}_{1:T}} p(F \mid \mathbf{x}_{1:T}) p(\mathbf{x}_{1:T} \mid c_0, \rho) \, \mathrm{d}\mathbf{x}_{1:T}, \tag{7.9}$$

where as before ρ is the unobserved property. This can now be approximated through simple Monte Carlo, where we roll out simulations $\mathbf{x}_{1:T}^{(i)}$ for $1 \leq i < N$, as described in Chapter 4. If simple Monte Carlo is our algorithm, then our hyperparameters are T (the number of timesteps) and N (the number of simulations). Additionally, as discussed above, we will also have any representational hyperparameters of the simulation itself, such as the level of detail of the simulation.

However, as I mentioned previously, the precise representation can be quite difficult to pin down, and so I do not analyze the representational hyperparameters further here.

Based on Equation 7.9 and the identification of N and T as the hyperparameters, it should now be clear what questions need to be answered. First, how long should the simulations be run for (T), and second, how many simulations should be run (N)? Additionally, if ρ exists in a continuous and/or high dimensional space, then we can additionally ask: which ρ should we sample?

7.2.5.2 Sequential decision making

Sequential decision making problems are much more difficult to solve than basic inference or prediction problems because they involve taking actions which affect the state of the world and the rewards that are received, often delayed from the actions that generated them. Even Markov Decision Problems (MDPs), which are simpler versions of POMDPs, are notoriously difficult to solve and there are multiple subfields of computer science (such as planning and reinforcement learning) concerned with just that. Thus, there are many candidates for approximate algorithms that solve MDPs, and likely many more that have yet to be discovered. To give just one example as a case study, I focus on the particular algorithm of Monte-Carlo Tree Search (MCTS) (Coulom, 2006), which has been used successfully to beat human champions at the game of Go (Silver et al., 2016).

MCTS works by exploring a search tree using what is known as the *tree policy* and then estimating the value of nodes in that tree using Monte Carlo simulations generated via a *default policy* (or, alternatively, *rollout policy*). We can think of the default policy as being the mental simulation component of the algorithm, and the tree policy as being a meta-level policy which determines which parts of the state space to explore. Different variants of MCTS use different tree policies, but they usually work by selecting states to trade-off between maximizing the Q-function (exploitation) and visiting states that have not previously been visited (exploration). For example, one common tree policy is to use the Upper Confidence Bound for Trees (UCT) algorithm (Kocsis & Szepesvári, 2006), which treats the choice of actions as a multi-armed bandit problem. However, this approach results in suboptimal behavior because the choice of actions via the tree policy does not result in immediate, real rewards as is the assumption in bandit problems (Hay et al., 2012).

A more accurate view of the tree policy is that it is a way to *gather evidence* about which actions will be most valuable. It does not actually matter what rewards we receive in simulation by following the tree policy, as long as we can choose the actions that results in the highest reward at test time. However, what *does* matter is the amount of time we take exploring the tree. Thus, what the tree policy should maximize is a combination of how useful the computations will be with how time consuming they will be. This formulation begins to look a lot like metareasoning (Russell & Wefald, 1991), where the "hyperparameters" to be optimized are the choice of which states to explore. An additional hyperparameter that could be optimized here is the length of the rollout policy, corresponding to T in the earlier examples.

Indeed, Hay et al. (2012) applied metareasoning to the tree policy of MCTS and found it outperformed standard bandit-based methods. The challenge is in applying metareasoning to arbitrary-sized trees and for more challenging state and action spaces (such as high-dimensional, continuous spaces). While this is an ongoing area of research, recent work in deep learning has suggested promising methods for approximating the meta-level policy dictated by metareasoning (Hamrick et al., 2017; Pascanu et al., 2017).

The power of imagination makes us infinite.

John Muir

8 Conclusion

In this thesis, I have presented a new way of thinking about mental simulation: that it is a computational tool to be wielded in tandem with other cognitive algorithms and strategies. I have argued that to have a deep understanding of mental simulation—to understand not just *what* it is, but also *how* it is used—it is necessary to move beyond debates about whether or not mental simulation exists and to ask (1) under what circumstances is mental simulation the appropriate tool for the job, and (2) what is the most effective way to make use of it? This thesis is the first attempt to answer these questions both behaviorally and quantitatively, providing models not just of simulation itself but also how meta-level decisions are made about how to use simulation.

8.1 Summary

In Chapter 2 ("Background") and Chapter 3 ("Formalizing Mental Simulation"), I reviewed the vast literature on mental simulation and argued that the one thing that unites everything referred to as "mental simulation" is the use of a *transition function* or *observation function*. In other words, if the reasoning process involves some sort of model of how things ought to change over time or as a result of actions (and if that model is actively used to make predictions and inferences), then it is mental simulation. Framing the phenomena of mental simulation in this way allows us to ask more precise questions about its nature: what are the representations used in the transition and observation functions, how long should simulations be run for, how many simulations should be run, and which simulations should be chosen?

In Chapter 4 ("Learning by Thinking"), I dove more deeply into questions surrounding how we actually learn from mental simulation. In particular, given the capacity to run a mental simulation, what should be done with the results? For example, given a simulation of physical dynamics of a tower of blocks, how should one conclude whether the tower fell or which color block is heavier? While the answer may seem intuitively obvious—it feels as if one just "knows" the answer—specifying quantitatively where that answer comes from is less so. I proposed that mental simulations can be used in conjunction with Bayesian inference in order to make rich inferences about the world around us. I presented three experiments that demonstrated people's capacity to reason about the relative masses of objects in naturalistic 3D scenes. I found that people make accurate inferences, and that they continue to fine-tune their beliefs over time. To explain the results, I proposed a cognitive model that combines Bayesian inference with approximate knowledge of Newtonian physics by estimating probabilities from noisy physical simulations. I found that this model accurately predicts judgments from our experiments, suggesting that the same simulation mechanism underlies both peoples' predictions and inferences about the physical world around them.

In Chapter 5 ("Selecting Computations"), I focused on the question of how people decide what simulations to run. To answer this question, I explored how mental simulation should be used in the classic psychological task of determining if two images depict the same object in different orientations (Shepard & Metzler, 1971). Through a rational analysis of mental rotation, I formalized four models and compared them to human performance. I found that three models based on previous hypotheses in the literature were unable to account for several aspects of human behavior. The fourth was based on the idea of active sampling (e.g. Gureckis & Markant, 2012), which is a strategy of choosing actions that will provide the most information. This last model provided a plausible account of how people choose mental rotations, where the other models do not. Based on these results, I suggested that the question of "what to simulate?" is more difficult than has previously been assumed, and that an active learning approach holds promise for uncovering the answer.

Chapter 6 ("Allocation of Cognitive Resources") was concerned with the question of how the mind should allocate its cognitive resources. In this chapter, I investigated how people allocate resources for mental simulations, focusing in particular on the question of whether people vary the number of simulations that they run in order to optimally balance speed and accuracy. I combined a model of noisy physical simulation with a decision making strategy called the sequential probability ratio test, or the SPRT (Wald, 1947), which accumulates evidence until a threshold has been reached. The model predicted that people should use more samples when it is harder to make an accurate prediction due to higher simulation uncertainty, and that they should adjust their evidence threshold based on the payoff structure of the world. I tested these predictions across three experiments, each involving a task in which people had to judge whether a ball bouncing in a box would go through a hole or not. Across experiments, I varied the payoff structure of the task, and within each experiment I varied the uncertainty across trials by changing the size of the holes and the margin by which the ball went through or missed the hole. Both people's judgments and response times were well-predicted by the model, demonstrating that people have a systematic strategy to allocate resources for mental simulation.

Finally, in Chapter 7 ("A Formal Framework for Modeling Mental Simulation"), I introduced a framework for thinking about and modeling mental simulation. I proposed that all problems for which mental simulation is a solution can be formulated at the computational level of analysis (Marr, 1982) as a partially observable Markov decision problem, or POMDP (Kaelbling et al., 1998). I argued that while such problems do not necessarily need to be solved by mental simulation, they often are because mental simulation is general and flexible enough to cover many situations. However, to understand in more detail why and how mental simulation is used, it is necessary to attend to the algorithmic level of analysis. Specifically, I argued for the use of resource-rational analysis (Griffiths et al., 2015) applied to simulation-based algorithms in order to generate hypotheses about how people answer the questions of what simulations to run, how many to run, how long to run them for, and so on. By viewing mental simulation through the lenses of POMDPs and resource-rational analysis, I believe we will gain not only a deeper understanding of how the mind manages its computational resources, but also the mechanisms which underly its breathtaking flexibility and capacity for generalization.

8.2 Looking Forward

This thesis is only the beginning of what I believe will be a fruitful and exciting area of research. Indeed, the research presented here has only just begun to touch on the core questions surrounding metareasoning and mental simulation. There are many questions yet unanswered, and I therefore conclude with some final thoughts and speculations regarding the questions I find most exciting and ripe for exploration.

How do people determine the length of simulations to run? In Chapters 5 and 6, I investigated the meta-level questions of which simulations will be most useful to run, and how many should be run? The most natural next question is: how long should each individual simulation be run? The answer depends critically on the task being performed; for example, in determining whether a tower of blocks will fall, only a short simulation needs to be run (to determine if any blocks are moving), while determining how far the blocks will scatter requires a much longer simulation (Battaglia et al., 2013). But, it is not clear how people know the length of the simulation to run *a priori*. There are a few possibilities how how the length might be determined. One is that people learn from experience approximately how long simulations need to be for a given simulation, and simply guess at the length. Another possibility is that people monitor their simulations and interrupt them when the information being gained from the simulation is no longer useful. Both of these possibilities imply different representational-level algorithms (Marr, 1982) and can likely be investigated through the methods I discussed in Chapter 7.

Are there intermediate representations between qualitative and detailed simulation, and how do we pick the appropriate level of detail? As discussed at length in Chapter 2, many cognitive phenomena have been attributed to some process of mental simulation. However, such mental simulations can vary widely in their level of detail, sometimes comprising more abstract (Khemlani et al., 2013) or "qualitative" simulations (Forbus, 1983), while other times seeming to be highly detailed (Battaglia et al., 2013; Smith & Vul, 2013). This variation in the granularity of mental simulation likely reflects a more general capacity for people to choose the right level of representation beyond just the two levels of detail that have typically been discussed. Indeed, we may have the ability to access a whole spectrum of representations for simulation, ranging from coarse to highly detailed as demanded by the task. It is an interesting question to ask: how broad is this range, and how appropriate (or efficient) is the choice of representation? As with the question regarding how long simulations ought to be run for, I believe these questions regarding the detail of simulation can also be investigated through the methods discussed in Chapter 7.

Can people learn alternate strategies from simulation, or vice versa? I discussed in Chapter 7 the likelihood that people have multiple strategies that they switch between. Yet, this raises the question: where do those alternate strategies come from? One possibility is that people may learn specialized rules or heuristics from mental simulations (Callaway, Hamrick, & Griffiths, 2017; Schwartz & Black, 1996). Another possibility is that we pick up a collection of rules, heuristics, and isolated memories through experience and gradually stitch them together into a more general capacity for mental simulation (e.g. Barsalou, 1999). Both of these possibilities are likely to have some truth to them, but it remains to be seen exactly how these two learning processes interact.

What is the relationship between mental simulation and creativity? There is a sense in which mental simulation is not only useful for prediction and inference, but also for *exploration*. For example, Finke and Slayton (1988) showed that people can use mental simulation to discover novel and creative combinations of simple objects, such as combining the shapes V, O, and D into an ice cream cone.[1] Given mental simulation's importance in planning, which inherently must deal with the exploration-exploitation trade-off (Sutton & Barto, 1998), it seems unsurprising that the mind might be adapted towards using mental simulation for exploratory thought more generally. Moreover, although the way we choose which mental simulations to run is biased towards what is normally expected, we are able to modify our simulations to explore just outside the bounds of everyday life. For example, I can imagine riding an animal (which is not too far out of the ordinary), but gradually change the features of my mental simulation to make it unordinary (such as imagining riding a giant cat, or a flying alligator). The fact that our mental simulations tend to encode so much that is true about the world (Gendler, 1998) may be what makes them so perfect for creative thought: by tweaking them just by a small amount, we hit the sweet spot between novelty and utility that seems crucial for marking something as "creative" (Ward, 1994). Of course, many of these ideas have been said before (Ward, Smith, & Finke, 1999). However, now that we are beginning to have better models of mental simulation and of the meta-level decisions surrounding mental simulation, it is likely that we will be able to explore these ideas even further.

How do people use their mental simulations of other agents? In this dissertation, I have focused on the domain of physical reasoning because we know the ground-truth physical dynamics in the world. However, as outlined in Chapter 2, mental simulation is used not just in physical scenes but also in social scenes: we imagine what other agents believe, what they desire, and what actions they might take. Modeling this "theory of mind" has recently been a focus both of cognitive scientists (e.g. Baker et al., 2009; Baker & Tenenbaum, 2014; Evans, Stuhl, Goodman, Stuhlmüller, & Goodman, 2016; Jara-Ettinger et al., 2016; Kleiman-Weiner, Ho, Austerweil, Littman, & Tenenbaum, 2016; Ullman et al., 2009) as well as computer scientists interested in human-robot interaction (e.g. Dragan, Bauman, Forlizzi, & Srinivasa, 2015; Dragan, Lee, &

[1] Rotate the D by 90° left, place it on the V, and then place the O on the very top.

Srinivasa, 2013; Fisac et al., 2017, 2016; Huang, Held, Abbeel, & Dragan, 2017; Liu et al., 2016; Sadigh, Dragan, Sastry, & Seshia, 2017; Sadigh, Sastry, Seshia, & Dragan, 2016). While we are beginning to have better and better models of human theory of mind, most of the research on theory of mind focuses on characterizing the nature of the simulation itself, much in the same way that most research on mental imagery focuses on characterizing the nature of the imagery. What is often missing are the meta-level questions surrounding *how* our models of other agents are used. For example, it is not required to consider all aspects of your partner's personality in order to predict whether they will be pleasantly surprised by a gift. How do we selectively choose which parts of our theory of mind to simulate? Similarly, our actions might affect others on different timescales, from seconds to years. How do we choose the right level of abstraction for running these simulations? As our models of theory of mind continue to improve, we can hopefully begin to model the answers to such meta-level questions as well.

How can mental simulation inform artificial intelligence? The newly re-awakened field of artificial intelligence has generated much attention and excitement, with impressive advances in areas such as object recognition (Krizhevsky et al., 2012), game playing (Mnih et al., 2015; Silver et al., 2016), and robotic control (Levine, Finn, Darrell, & Abbeel, 2016; Lillicrap et al., 2016). Yet, no AI system today can hold a candle to the flexibility with which people navigate the world. Human cognition owes this achievement, in part, to its ability to simulate what will happen in arbitrarily complex and novel situations: as I have argued in this thesis, people automatically know what simulations to run, for how long, at what level of detail, and how to interpret the results. These ideas have started to make their way in to AI research under the term "imagination" (Hamrick et al., 2017; Pascanu et al., 2017; Weber et al., 2017), and researchers both from cognitive science and neuroscience have called on the field of AI to pay more attention to human cognition (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Lake, Ullman, Tenenbaum, & Gershman, in press). By studying how humans engage in metareasoning and mental simulation, I believe we will gain key insights into how we can build similarly flexible artificial agents. Additionally, while AI need not mimic human intelligence, it *does* need to interact and communicate with people. It is important that AI is built with the quirks of human intelligence in mind, which requires that researchers in computer science work together with researchers in cognitive science, psychology, and other fields in the social sciences.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from http://tensorflow.org/ (Software available from tensorflow.org)

Aleman, A., Nieuwenstein, M. R., Böcker, K. B., & de Haan, E. H. (2000). Music training and mental imagery ability. *Neuropsychologia*, *38*(12), 1664–1668. doi: 10.1016/S0028-3932(00)00079-8

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (2015). Problem solving. In *Cognitive psychology and its implications* (8th ed., pp. 181–209). New York, NY: Worth Publishers.

Arditi, A., Holtzman, J. D., & Kosslyn, S. M. (1988). Mental imagery and sensory experience in congenital blindness. *Neuropsychologia*, *26*(1), 1–12. doi: 10.1016/0028-3932(88)90026-7

Aristotle. (350 BCE). *On The Soul* (J. A. Smith, Trans.). The Internet Classics Archive, MIT.

Atkeson, C., & Santamaria, J. (1997). A comparison of direct and model-based reinforcement learning. *Proceedings of International Conference on Robotics and Automation*, 3557–3564. doi: 10.1109/ROBOT.1997.606886

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009, dec). Action understanding as inverse planning. *Cognition*, *113*(3), 329–49. doi: 10.1016/j.cognition.2009.07.005

Baker, C. L., & Tenenbaum, J. B. (2014). Modeling Human Plan Recognition using Bayesian Theory of Mind. In *Plan, activity, and intent recognition* (pp. 177–204). doi: 10.1016/B978-0-12-398532-3.00007-5

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences*, *22*(4), 577–609; discussion 610–60.

Bartlett, F. C. (1932). *Remembering: An experimental and social study.*

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. *CogSci 2015*, 172–177.

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

Baugh, L. A., Kao, M., Johansson, R. S., & Flanagan, J. R. (2012, September). Material evidence: interaction of well-learned priors and sensorimotor memory when lifting objects. *Journal of Neurophysiology*, *108*(5), 1262–9.

Bellman, R. E. (1957). *Dynamic Programming.* Princeton, NJ: Princeton University Press.

Belousov, B., Neumann, G., Rothkopf, C. A., & Peters, J. R. (2016). Catching heuristics are optimal control policies. *Advances in Neural Information Processing Systems 29 (NIPS 2016).*

Bensafi, M., Porter, J., Pouliot, S., Mainland, J., Johnson, B., Zelano, C., ... Sobel, N. (2003). Olfactomotor activity during imagery mimics that during perception. *Nature Neuroscience*, *6*(11), 1142–1144. doi: 10.1038/nn1145

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007, sep). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733–64. doi: 10.1080/03640210701530748

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*(102), 1–17.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. doi: 10.1080/01621459.2017.1285773

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. doi: 10.1037/0033-295X.113.4.700

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classiiers. *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT 1992).*

Bramley, N. R., Gerstenberg, T., & Tenenbaum, J. B. (2016). Natural science: Active learning in dynamic physical microworlds. In *Proceedings of the 38th annual meeting of the cognitive science society.*

Buckingham, G., Ranger, N. S., & Goodale, M. A. (2011). The material-weight illusion induced by expectations alone. *Attention, Perception & Psychophysics*, *73*(1), 36–41. doi: 10.3758/s13414-010-0007-4

Bullet Developers. (2013). *Bullet Collision Detection and Physics Library.* http://www.bulletphysics.org/.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 214–220. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1202143 doi: 10.1037/0096-1523.1.3.214

Callaway, F., Hamrick, J. B., & Griffiths, T. L. (2017). Discovering simple heuristics from mental simulation. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

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

Charpentier, A. (1891). Analyse experimentale quelques elements de la sensation de poids [Experimental study of some aspects of weight perception]. *Archives de Physiologie Normales Pathologiques*, *3*, 122–135.

Clement, J. J. (2009, oct). The Role of Imagistic Simulation in Scientific Thought Experiments. *Topics in Cognitive Science*, *1*(4), 686–710. doi: 10.1111/j.1756-8765.2009.01031.x

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, *7*, 20–43.

Coulom, R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. In *Proceedings of the 5th international conference on computer and games.*

Craik, K. (1943). Hypothesis on the nature of thought. In *The nature of explanation* (pp. 50–61).

Davis, E. (2008). Pouring liquids: A study in commonsense physical reasoning. *Artificial Intelligence*, *172*(12-13), 1540–1578. doi: 10.1016/j.artint.2008.04.003

Davis, E. (2010). How does a box work? A study in the qualitative dynamics of solid objects. *Artificial Intelligence*, *175*(1), 299–345. doi: 10.1016/j.artint.2010.04.006

Davis, E., Marcus, G., & Chen, A. (2013). Reasoning from Radically Incomplete Information: The Case of Containers. *Advances in Cognitive Systems*, *2*, 1–18.

Davis, E., & Marcus, G. F. (2014). The Scope and Limits of Simulation in Cognitive Models. *arXiv:1506.04956 [cs.AI]*, 1–27.

de Kleer, J., & Brown, J. S. (1984). A Qualitative Physics Confluences. *Artificial Intelligence*, *24*(December), 7–83. doi: 10.1016/0004-3702(84)90037-7

Denil, M., Agrawal, P., Kulkarni, T. D., Erez, T., Battaglia, P. W., & de Freitas, N. (2017). Learning to Perform Physics Experiments via Deep Reinforcement Learning. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, *1*, 163–175.

Dils, A. T., & Boroditsky, L. (2010, sep). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16396–16400. doi: 10.1073/pnas.1009438107

Dragan, A. D., Bauman, S., Forlizzi, J., & Srinivasa, S. S. (2015). Effects of Robot Motion on Human-Robot Collaboration. In *Proceedings of the 10th annual acm/ieee international conference on human-robot interaction.*

Dragan, A. D., Lee, K., & Srinivasa, S. (2013). Legibility and predictability of robot motion. *International Conference on Human-Robot Interaction*. doi: 10.1109/HRI.2013.6483603

Duane, S., Kennedy, A., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. doi: 10.1016/0370-2693(87)91197-X

Duvenaud, D. (2013). *Log-gaussian processes for Bayesian quadrature.* Personal communication.

Ellis, R. R., & Lederman, S. J. (1998). The golf-ball illusion: evidence for top-down processing in weight perception. *Perception*, *27*(1917), 193–201. doi: 10.1068/p270193

Ellis, R. R., & Lederman, S. J. (1999). The material-weight illusion revisited. *Perception & Psychophysics*, *61*(8), 1564–1576. doi: 10.3758/BF03213118

Evans, O., Stuhl, A., Goodman, N. D., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the Preferences of Ignorant, Inconsistent Agents. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 323–329.

Farah, M. J., Hammond, K. M., Levine, D. N., & Calvanio, R. (1988). Visual and spatial mental imagery: Dissociable systems of representation. *Cognitive Psychology*, *20*(4), 439–462. doi: 10.1016/0010-0285(88)90012-6

Faw, B. (2009). Conflicting intuitions may be based on differing abilities: Evidence from mental imaging research. *Journal of Consciousness Studies*, *16*(4), 45–68.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3rd ed., Vol. 1). John Wiley & Sons Incorporated.

Feltz, D. L., & Landers, D. M. (1983). The Effects of Mental Practice on Motor Skill Learning and Performance: A Meta-analysis. *Journal of Sport and Exercise Psychology*, *5*(1), 25–57.

Finke, R. A., & Slayton, K. (1988). Explorations of creative visual synthesis in mental imagery. *Memory & Cognition*, *16*(3), 252–257. doi: 10.3758/BF03197758

Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., … Dragan, A. D. (2017). Pragmatic-Pedagogic Value Alignment. *arXiv preprint arXiv:1707.06354v1*.

Fisac, J. F., Liu, C., Hamrick, J. B., Sastry, S., Hedrick, J. K., Griffiths, T. L., & Dragan, A. D. (2016). Generating Plans that Predict Themselves. In *Proceedings of the 12th international workshop on the algorithmic foundations of robotics (wafr 2016).*

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). The functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1610344113

Flanagan, J. R., & Beltzner, M. A. (2000). Independence of perceptual and sensorimotor predictions in the size-weight illusion. *Nature Neuroscience*, *3*(7), 737–741. doi: 10.1038/76701

Flanagan, J. R., Bittner, J. P., & Johansson, R. S. (2008). Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current Biology*, *18*(22), 1742–1747. doi: 10.1016/j.cub.2008.09.042

Flanagan, J. R., Vetter, P., Johansson, R. S., & Wolpert, D. M. (2003). Prediction precedes control in motor learning. *Current Biology*, *13*(2), 146–150. doi: 10.1016/S0960-9822(03)00007-1

Forbus, K. D. (1983). Qualitative Reasoning About Space and Motion. In *Mental models* (pp. 53–73).

Forbus, K. D. (2011, jul). Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(4), 374–391. doi: 10.1002/wcs.115

Frank, M. C., & Barner, D. (2011). Representing exact number visually using mental abacus. *Journal of Experimental Psychology: General*, *141*, 134–149.

Freyd, J. J., & Finke, R. A. (1984). Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126–132. doi: 10.1037/0278-7393.10.1.126

Freyd, J. J., & Jones, K. T. (1994). Representational Momentum for a Spiral Path. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 968–976. doi: 10.1037/0278-7393.20.4.968

Freyd, J. J., Pantzer, T. M., & Cheng, J. L. (1988). Representing Statics as Forces in Equilibrium. *Journal of Experimental Psychology: General*, *117*, 395–407. doi: dx.doi.org/10.1037/0096-3445.117.4.395

Funt, B. V. (1983). A parallel-process model of mental rotation. *Cognitive Science*, *7*(1), 67–93.

Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501. doi: 10.1016/S1364-6613(98)01262-5

Galton, F. (1880). Statistics of Mental Imagery. *Mind*, *5*, 301–318.

Gardin, F., & Meltzer, B. (1989). Analogical Representation of Naive Physics. *Artifical Intelligence*, *38*(1989), 139–159.

Gardner, H. (1987). Cognitive Science: The First Decades. *The Mind's New Science. A History of the Cognitive Revolution*(1958), 430.

Gardony, A. L., Taylor, H. A., & Brunye, T. T. (2014). What does physical rotation reveal about mental rotation? *Psychological Science*, *25*(2), 605–612.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. doi: 10.1109/TPAMI.1984.4767596

Gendler, T. S. (1998). Galileo and the Indispensability of Scientific Thought Experiment. *The British Journal for the Philosophy of Science*, *49*(3), 397–424.

Gentner, D., & Stevens, A. (Eds.). (1983). *Mental Models*. Lawrence Erlbaum Associates.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi: 10.1126/science.aac6076

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.*

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Gerstenberg, T., Peterson, M. F., Goodman, N. D., & Tenenbaum, J. B. (2017). Eye-tracking causality. *under review*.

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty Towers: A counterfactual simulation model of physical support. In *Proceedings of the 39th annual conference of the cognitive science society.*

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

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

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

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

Glasgow, J. I., & Papadias, D. (1992). Computational imagery. *Cognitive Science*, *16*(3), 355–394.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*, 535–574.

Goldman, A. I. (1992). In Defense of the Simulation Theory. *Mind & Language*, *7*(1-2), 104–119. doi: 10.1111/j.1468-0017.1992.tb00200.x

Gopnik, A., & Wellman, H. M. (1992). Why the Child's Theory of Mind Really Is a Theory. *Mind & Language*, *7*(1-2), 145–171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Gordon, R. M. (1992). The Simulation theory: Objections and misconceptions. *Mind & Language*, *7*(1-2), 11–34. doi: 10.1111/j.1468-0017.1992.tb00195.x

Grandy, M. S., & Westwood, D. A. (2006). Opposite Perceptual and Sensorimotor Responses to a Size-Weight Illusion. *Journal of Neurophysiology*, *95*(6), 3887–3892. doi: 10.1152/jn.00851.2005

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., … Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N., & Battaglia, P. W. (2017). Metacontrol for adaptive imagination-based optimization. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex physical scenes via probabilistic simulation. *Cognition*, *157*, 61–76. doi: 10.1016/j.cognition.2016.08.012

Hamrick, J. B., Battaglia, P. W., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1–6).

Hamrick, J. B., & Griffiths, T. L. (2014). What to simulate? Inferring the right direction for mental rotation. *Annual Conference of the Cognitive Science Society*, 577–582.

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 *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1–6).

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. Retrieved from http://dx.doi.org/10.1016/j.neuron.2017.06.011 doi: 10.1016/j.neuron.2017.06.011

Hassabis, D., & Maguire, E. A. (2009). The Construction System of the Brain. *Philosophical Transactions of the Royal Soceity B*, *364*, 1263–1271. doi: 10.1098/rstb.2008.0296

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97–109.

Haueisen, J., & Knösche, T. R. (2001). Involuntary motor activity in pianists evoked by music perception. *Journal of cognitive neuroscience*, *13*(6), 786–92. doi: 10.1162/08989290152541449

Hay, N. J., Russell, S. J., Tolpin, D., & Shimony, S. E. (2012). Selecting Computations: Theory and Applications. *arXiv preprint arXiv:1207.5878v1 [cs.AI]*.

Hayes, P. J. (1979). The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the micro-electronic age* (pp. 242–270). Edinburgh University Press.

Hayes, P. J. (1985). *The Second Naive Physics Manifesto* (J. Hobbs & B. Moore, Eds.). Ablex Publishing Corporation.

Hegarty, M. (1992). Mental Animation: Inferring Motion From Static Displays of Mechanical Systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1084–1102.

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

Helmholtz, H. (1925). Concerning the perceptions in general. In J. P. C. Southall (Ed.), *Treatise on physiological optics* (pp. 1–37). New York: Dover Publications.

Hollins, M. (1985). Styles of mental imagery in blind adults. *Neuropsychologia*, *23*(4), 561–566. doi: 10.1016/0028-3932(85)90009-0

Huang, S. H., Held, D., Abbeel, P., & Dragan, A. D. (2017). Enabling Robots to Communicate their Objectives. *Robotics: Science and Systems*. Retrieved from http://arxiv.org/abs/1702.03465

Hubbard, T. L. (1996). Representational momentum, centripetal force, and curvilinear impetus. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 1049–1060. doi: 10.1037/0278-7393.22.4.1049

Hubbard, T. L. (1997). Target size and displacement along the axis of implied gravitational attraction: Effects of implied weight and evidence of representational gravity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1484–1493. doi: 10.1037/0278-7393.23.6.1484

Hubbard, T. L. (2004). The Perception of Causality: Insights from Michotte's Launching Effect, Naïve Impetus Theory, and Representational Momentum. In A. M. Oliveira, M. Teixeira, G. F. Borges, & M. J. Ferro (Eds.), *Fechner Day* (pp. 116–121). Coimbra, Portugal: The International Society for Psychophysics.

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*(5), 822–851. doi: 10.3758/BF03196775

Hubbard, T. L. (2013a). Launching, Entraining, and Representational Momentum: Evidence Consistent with an Impetus Heuristic in Perception of Causality. *Axiomathes*, *23*(4), 633–643. doi: 10.1007/s10516-012-9186-z

Hubbard, T. L. (2013b). Phenomenal Causality II: Integration and Implication. *Axiomathes*, *23*, 485–524. doi: 10.1007/s10516-012-9200-5

Hubbard, T. L. (2013c). Phenomenal Causality I: Varieties and Variables. *Axiomathes*, *23*, 1–42. doi: 10.1007/s10516-012-9198-8

Hubbard, T. L., & Ruppel, S. E. (2002). A possible role of naïve impetus in Michotte's "launching effect": Evidence from representational momentum. *Visual Cognition*, *9*(1-2), 153–176. doi: 10.1080/13506280143000377

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95.

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. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661316300535 doi: 10.1016/j.tics.2016.05.011

Jeannerod, M. (1995). Mental Imagery in the Motor Context. *Neuropsychologia*, *33*(11), 1419–1432.

Jennings, J., & Davies, J. (2017). The Motor System Does Not Use a Curvilinear Impetus Belief: Folk Physics and Embodied Cognition. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Johnson-Laird, P. N. (2012). Inference with Mental Models. In *The oxford handbook of thinking and reasoning* (pp. 134–145). doi: 10.1093/oxfordhb/9780199734689.001.0001

Johnson-Laird, P. N., & Yang, Y. (2008). Mental Logic, Mental Models, and Simulations of Human Deductive Reasoning. *The Cambridge Handbook of Computational Psychology*(1993), 339–358. doi: 10.1017/CBO9780511816772

Julstrom, B. A., & Baron, R. J. (1985). A model of mental imagery. *International Journal of Man-Machine Studies*, *23*, 313–334.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Don't stop 'til you get enough: adaptive information sampling in a visuomotor estimation task. In *Proceedings of the 33rd Annual Meeting of the Cogntive Science Society* (pp. 2854–2859).

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*, 137–172.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Kaiser, M. K., Proffitt, D. R., & McCloskey, M. (1985). The development of beliefs about falling objects. *Perception & Psychophysics*, *38*(6), 533–539. doi: 10.3758/BF03207062

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinions in Neurobiology*, *9*(6), 718–727. doi: 10.1016/S0959-4388(99)00028-8

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: retrieval as mental simulation. *Trends in Cognitive Sciences*, *12*(3), 92–98. doi: 10.1016/j.tics.2007.12.004

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(42), 16766–71. doi: 10.1073/pnas.1316275110

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. *Proceedings of the European Conference on Machine Learning*.

Kosslyn, S. M. (1988). Aspect of a Cognitive Neuroscience of Mental Imagery. *Science*, *240*(4859), 1621–1626.

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), 47–60. doi: 10.1037/0096-1523.4.1.47

Kosslyn, S. M., Pinker, S., Smith, G. E., & Shwartz, S. P. (1979). On the demystification of mental imagery. *The Behavioral and Brain Sciences*, *2*, 535–581.

Kosslyn, S. M., & Shwartz, S. P. (1977). A simulation of visual imagery. *Cognitive Science*, *1*, 265–295.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The Case for Mental Imagery*. Oxford University Press.

Kozhevnikov, M., & Hegarty, M. (2001, sep). Impetus beliefs as default heuristics: dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, *8*(3), 439–53.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information and Processing Systems*.

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 meeting of the cognitive science society.*

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 conference of the cognitive science society.*

Kuipers, B. (1986). Qualitative Simulation. *Artificial Intelligence*, *29*(3), 289–338. doi: 10.1016/0004-3702(86)90073-1

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (in press). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*.

Levillain, F., & Bonatti, L. L. (2011). A Dissociation Between Judged Causality and Imagined Locations in Simple Dynamic Scenes. *Psychological Science*, *22*(5), 674–681. doi: 10.1177/0956797611404086

Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-End Training of Deep Visuomotor Policies. *Journal of Machine Learning Research*, *17*, 1–40.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, *25*, 1–9.

Lieder, F., Hamrick, J. B., Hay, N. J., Plunkett, D., Russell, S. J., & Griffiths, T. L. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. *Advances in Neural Information Processing Systems*, *27*, 2870–2878.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Hess, N., Erez, T., Tassa, Y., … Wierstra, D. (2016). Continuous control with deep reinforcement learning. *Proceedings of the International Conference on Learning Representations (ICLR 2016).*

Liu, C., Hamrick, J. B., Fisac, J. F., Dragan, A. D., Hedrick, J. K., Sastry, S. S., & Griffiths, T. L. (2016). Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration. In J. Thangarajah, K. Tuyls, C. Jonker, & S. Marsella (Eds.), *Proceedings of the 15th international conference on autonomous agents and multiagent systems (aamas 2016).* Singapore.

Lombrozo, T. (in press). "Learning by thinking" in science and everyday life. In P. Godfrey-Smith & A. Levy (Eds.), *The scientific imagination.* Oxford University Press.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, *2*.

Lupo, J., & Barnett-Cowan, M. (2015). Perceived object stability depends on shape and material properties. *Vision Research*, *109*, 158–65. doi: 10.1016/j.visres.2014.11.004

Marr, D. (1982). The Philosophy and the Approach. In *Vision* (pp. 8–37).

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & Cognition*, *32*(8), 1389–1400. doi: 10.3758/BF03206329

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130. doi: 10.1038/scientificamerican0483-122

McCloskey, M., & Kohl, D. (1983). Naive physics: the curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 146–156. doi: 10.1037/0278-7393.9.1.146

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive Physics: The Straight-Down Belief and Its Origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649.

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 51 - 56).

McLeod, P., Reed, N., & Dienes, Z. (2006). The Generalized Optic Acceleration Cancellation Theory of Catching. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 139–148. doi: 10.1037/0096-1523.32.1.139

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics*, *21*(6), 1087–1092. doi: http://dx.doi.org/10.1063/1.1699114

Milli, S., Lieder, F., & Griffiths, T. L. (2017). When Does Bounded-Optimal Metareasoning Favor Few Cognitive Systems? *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. a., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. Retrieved from http://dx.doi.org/10.1038/nature14236 doi: 10.1038/nature14236

Murray, D. J., Ellis, R. R., Bandomir, C. A., & Ross, H. E. (1999). Charpentier (1891) on the size-weight illusion. *Perception & Psychophysics*, *61*(8), 1681–1685. doi: 10.3758/BF03213127

Neal, R. M. (2003). Slice Sampling. *The Annals of Statistics*, *31*(3), 705–767.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain monte carlo* (chap. 5). Chapman & Hall.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a Theory of Human Problem Solving. *Psychological Review*, *65*(3). doi: 10.1037/h0048495

Ng, A. Y., & Jordan, M. I. (2002). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, *14*.

Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., & Ghahramani, Z. (2012). Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems 25* (pp. 46–54).

Panda3D Developers. (2013). *Panda3D v1.9.0*. https://www.panda3d.org/.

Parsons, L. M. (1994, aug). Temporal and kinematic properties of motor behavior reflected in mentally simulated action. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 709–730. doi: 10.1037/0096-1523.20.4.709

Pascanu, R., Li, Y., Vinyals, O., Heess, N., Buesing, L., Racanière, S., … Battaglia, P. (2017). Learning model-based planning from scratch. *arXiv preprint arXiv: 1707.06170*. Retrieved from https://arxiv.org/abs/1707.06170

Pérez, F., & Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. Retrieved from http://ipython.org doi: 10.1109/MCSE.2007.53

Perky, C. W. (1910). An experimental study of imagination. *The American Journal of Psychology*, *21*(3), 422–452.

Peterson, C., & Anderson, J. R. (1987). A Mean Field Theory Learning Algorithm for Neural Networks. *Complex Systems*, *1*, 995–1019.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Science*, *25*, 157–238.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. doi: 10.1163/156856888X00122

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* Cambridge, MA: MIT Press.

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

Ross, B. (2016). *Aphantasia: How It Feels To Be Blind In Your Mind.* https://www.facebook.com/notes/blake-ross/aphantasia-how-it-feels-to-be-blind-in-your-mind/10156834777480504/.

Ross, H. E. (1969). When is a weight not illusory? *The Quarterly Journal of Experimental Psychology*, *21*(4), 346–355. doi: 10.1080/14640746908400230

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

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

Russell, S. J., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*(1-3), 361–395. doi: 10.1016/0004-3702(91)90015-C

Sadigh, D., Dragan, A. D., Sastry, S., & Seshia, S. A. (2017). Active Preference-Based Learning of Reward Functions. *Robotics: Science and Systems*. doi: 10.15607/RSS.2017.XIII.053

Sadigh, D., Sastry, S., Seshia, S. A., & Dragan, A. D. (2016). Information Gathering Actions over Human Internal State. *International Conference on Intelligent Robotics*, 1–8. Retrieved from http://people.eecs.berkeley.edu/$\sim$anca/papers/IROS16_active.pdf doi: 10.1109/IROS.2016.7759036

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology*, *5*(938), 1–7. doi: 10.3389/fpsyg.2014.00938

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2009). A Bayesian framework for modeling intuitive dynamics. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1–6).

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

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences*, *9*(4), 174–179. doi: 10.1016/j.tics.2005.01.012

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, *76*(4), 677–694. doi: 10.1016/j.neuron.2012.11.001

Schwartz, D. L. (1999, may). Physical imagery: kinematic versus dynamic models. *Cognitive Psychology*, *38*(3), 433–464. doi: 10.1006/cogp.1998.0702

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science*, *20*(4), 457–497. doi: 10.1016/S0364-0213(99)80012-3

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

Seashore, C. E. (1899). Some psychological statistics. 2. The material-weight illusion. *University of Iowa Studies in Psychology*.

Sekuler, R., & Nash, D. (1972). Speed of size scaling in human vision. *Psychonomic Science*, *27*(2), 93–94.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Cooper, L. A. (1982). *Mental Images and their Transformations*. MIT Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., … Kavukcuoglu, K. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7585), 484–489. Retrieved from http://dx.doi.org/10.1038/nature16961 doi: 10.1038/nature16961

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex*, *15*(10), 1570–1583. doi: 10.1093/cercor/bhi035

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

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Smith, K. A., Peres, F. D. A. B., Vul, E., & Tenenbaum, J. B. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

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

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Stevenson, R. J., & Case, T. I. (2005). Olfactory imagery: a review. *Psychonomic Bulletin & Review*, *12*(2), 1–21. doi: 10.3758/BF03196369

Stigler, J. W. (1984). ”Mental abacus”: The effect of abacus training on Chinese children’s mental calculation. *Cognitive Psychology*, *16*(2), 145–176. doi: 10.1016/0010-0285(84)90006-9

Strauss Marmor, G., & Zaback, L. A. (1976). Mental Rotation by the Blind: Does Mental Rotation Depend on Visual Imagery? *Journal of Experimental Psychology: Human Perception and Perjormance*, *2*(4), 51–521. doi: 10.1037/0096-1523.2.4.515

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, *3*, 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011, may). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–9. doi: 10.1126/science.1196404

Tenenbaum, J. B., & Griffiths, T. L. (2001). The Rational Basis of Representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1–6).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011, mar). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–85. doi: 10.1126/science.1192788

Thomas, N. J. (2016). Mental imagery. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/mental-imagery/.

Tiggemann, M., & Kemps, E. (2005). The phenomenology of food cravings: The role of mental imagery. *Appetite*, *45*(3), 305–313. doi: 10.1016/j.appet.2005.06.004

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

Tong, C., Wolpert, D. M., & Flanagan, J. R. (2002). Kinematics and dynamics are not represented independently in motor working memory: evidence from an interference study. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *22*(3), 1108–1113. doi: 22/3/1108[pii]

Trickett, S. B., & Trafton, J. G. (2007, sep). "What if...": The Use of Conceptual Simulations in Scientific Reasoning. *Cognitive Science*, *31*(5), 843–875. doi: 10.1080/03640210701530771

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. *Advances in Neural Information and Processing Systems*.

Ullman, T. D., Spelke, E., Battaglia, P. W., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*. Retrieved from http://dx.doi.org/10.1016/j.tics.2017.05.012 doi: 10.1016/j.tics.2017.05.012

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2014). Learning physics from dynamical scenes. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

van der Walt, S., Colbert, S., & Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, *13*(2), 22-30. doi: 10.1109/MCSE.2011.37

van Opheusden, B., Bnaya, Z., Galbiati, G., & Ma, W. J. (2016). Do People Think Like Computers? *Computers and Games*, *10068*, 212–224. doi: 10.1007/978-3-319-50935-820

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637. doi: 10.1111/cogs.12101

Wald, A. (1947). *Sequential analysis*. Wiley, New York.

Wald, A., & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, *19*, 326–339. doi: 10.1214/aoms/1177730197

Wald, A., & Wolfowitz, J. (1950). Bayes solutions of sequential decision problems. *The Annals of Mathematical Statistics*, *21*(1), 82–99.

Walker, C. M., & Gopnik, A. (2013). Causality and Imagination. In M. Taylor (Ed.), *The development of imagination* (pp. 342—-358). New York: Oxford University Press.

Ward, T. B. (1994). *Structured Imagination: the Role of Category Structure in Exemplar Generation* (Vol. 27) (No. 1). Retrieved from http://www.sciencedirect.com/science/article/pii/S0010028584710103 doi: 10.1006/cogp.1994.1010

Ward, T. B., Smith, S. M., & Finke, R. A. (1999). Creative Cognition. In *Handbook of creativity* (pp. 189–212). New York: Cambridge University Press. doi: 10.1080/02783199909553973

Waskom, M., Botvinnik, O., Hobson, P., Cole, J. B., Halchenko, Y., Hoyer, S., … Allan, D. (2014). Seaborn v0.5.0 [Computer software manual]. Retrieved from http://dx.doi.org/10.5281/zenodo.12710 doi: {10.5281/zenodo.12710}

Weber, T., Racanière, S., Reichert, D. P., Buesing, L., Guez, A., Rezende, D. J., … Wierstra, D. (2017). Imagination-Augmented Agents for Deep Reinforcement Learning. *arXiv preprint arXiv: 1707.06203*. Retrieved from http://arxiv.org/abs/1707.06203

White, K. D. (1978). Salivation: The Significance of Imagery in Its Voluntary Control. *Psychophysiology*, *15*(3), 196–203.

White, P. A. (2012). The experience of force: The role of haptic experience of forces in visual perception of object motion and interactions, mental simulation, and motion-related judgments. *Psychological Bulletin*, *138*(4), 589–615. doi: 10.1037/a0025587

Willems, R. M., Toni, I., Hagoort, P., & Casasanto, D. (2010, oct). Neural dissociations between action verb understanding and motor imagery. *Journal of Cognitive Neuroscience*, *22*(10), 2387–400. doi: 10.1162/jocn.2009.21386

Winawer, J., Huk, A. C., & Boroditsky, L. (2008, mar). A motion aftereffect from still photographs depicting motion. *Psychological Science*, *19*(3), 276–283. doi: 10.1111/j.1467-9280.2008.02080.x

Wolfe, H. K. (1898). Some effects of size on judgments of weight. *Psychological Review*, *5*(1), 25–54. doi: 10.1037/h0073342

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. doi: 10.1016/S0960-9822(01)00432-8

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

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. B. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. In *Proceedings of the 39th annual conference of the cognitive science society.*

Yuille, J. C., & Steiger, J. H. (1982). Nonholistic processing in mental rotation: some suggestive evidence. *Perception & Psychophysics*, *31*(3), 201–209.

Zago, M., & Lacquaniti, F. (2005, sep). Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. *Journal of Neural Engineering*, *2*(3), S198–208. doi: 10.1088/1741-2560/2/3/S04

Zatorre, R. J., Halpern, A. R., & Ha, C. (2005). Mental Concerts: Musical Imagery and Auditory Cortex. *Neuron*, *47*, 9–12. doi: 10.1016/j.neuron.2005.06.013

Zeman, A., Della Sala, S., Torrens, L. A., Gountouna, V. E., McGonigle, D. J., & Logie, R. H. (2010). Loss of imagery phenomenology with intact visuo-spatial task performance: A case of 'blind imagination'. *Neuropsychologia*, *48*(1), 145–155. doi: 10.1016/j.neuropsychologia.2009.08.024

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery - Congenital aphantasia. *Cortex*, *73*, 378–380. doi: 10.1016/j.cortex.2015.05.019

Zwaan, R. A. (1999). Situation Models: The Mental Leap Into Imagined Worlds. *Current Directions in Psychological Science*, *8*(1), 15–18.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation Models in Language Comprehension and Memory. *Psychological Bulletin*, *123*(2), 162–185.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science*, *13*(2), 168–171.

A
Chapter 3 Details

A.1 Tower Stimuli

A.1.1 Construction

Subjects were presented with virtual, 3D tower scenes such as those in Figure 4.1b-c, which contained 10 rectangular blocks each with dimensions of 20cm×20cm×60cm. The blocks were stacked within a 120cm×120cm square column by a sequential random process such that when placed on the tower, no single block would fall off of its support (although, when later blocks were added, they might cause previously placed blocks to fall). This construction is the same as that used by Battaglia et al. (2013).

A.1.2 Selection

To choose the tower stimuli, we began with a set of 270 randomly generated geometries as well as the 30 tower geometries used in Battaglia et al. (2013). From each of these base towers, we generated 20 different versions in each of which 5 blocks were randomly assigned a label of "A" and 5 blocks were randomly assigned a label of "B". We ran model simulations for each of these 6000 towers under two mass ratios, $\kappa = 0.1$ and $\kappa = 10$, where the mass ratio is defined as the the ratio between the masses of "A" and "B" blocks.

For each stimulus $S_{i,j}$ (where $S_{i,j}$ is the j^{th} version of the i^{th} base tower), we computed whether the tower fell (F) under the true mass ratio. Then, according to Equation 4.4, we computed for both $\kappa = 0.1$ and $\kappa = 10$:

$$p(F|S_{i,j}, \kappa = k) \tag{A.1}$$

$$p(F|S_{i,j}, \kappa \neq k). \tag{A.2}$$

Equation A.1 is the likelihood given the *correct* hypothesis, while Equation A.2 is the likelihood given the *incorrect* hypothesis. If Equation A.2 is greater than Equation A.1 for a particular $S_{i,j}$, then our model will believe the *wrong* hypothesis. Based on earlier pilot data, people appear to be sensitive to this effect; thus, we excluded towers from consideration if Equation A.2 was greater than Equation A.1. Finally, we chose towers that maximized the likelihood ratio

$$\mathrm{LHR}(S_{i,j}, k) := \frac{p(F|S_{i,j}, \kappa = k)}{p(F|S_{i,j}, \kappa \neq k)} \tag{A.3}$$

for both $\kappa = 0.1$ and $\kappa = 10$. To do this, we found the best version S_i^* of each base stimulus:

$$S_i^* = \arg\max_{S_{i,j}} \mathrm{LHR}(S_{i,j}, \kappa = 0.1) \cdot \mathrm{LHR}(S_{i,j}, \kappa = 10) \tag{A.4}$$

and then chose the top 20 of those.

A.1.3 Rendering

Videos of the towers were rendered using Panda3D Developers (2013). In the training and prediction phases of Experiment 4.1, we rendered towers with red and blue blocks. In the inference phase of Experiment 4.1, we rendered towers with the following pairs of colors: blue-green, orange-blue, blue-yellow, gray-cyan, cyan-magenta, orange-cyan, cyan-purple, red-cyan, cyan-yellow, green-gray, gray-magenta, purple-gray, gray-red, magenta-green, green-orange, purple-green, magenta-yellow, orange-purple, yellow-purple, yellow-red. In the inference phases of Experiments 4.2-4.3, the towers were always rendered with purple and green blocks.

A.2 Fitting IPE Model Parameters

To obtain estimates from the IPE model, we ran 100 model samples for each stimulus for each of the parameter settings of $\sigma \in \{0.0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$ and $\phi \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.6, 2.0, 2.5, 3.0\}$, where σ represents perceptual un-

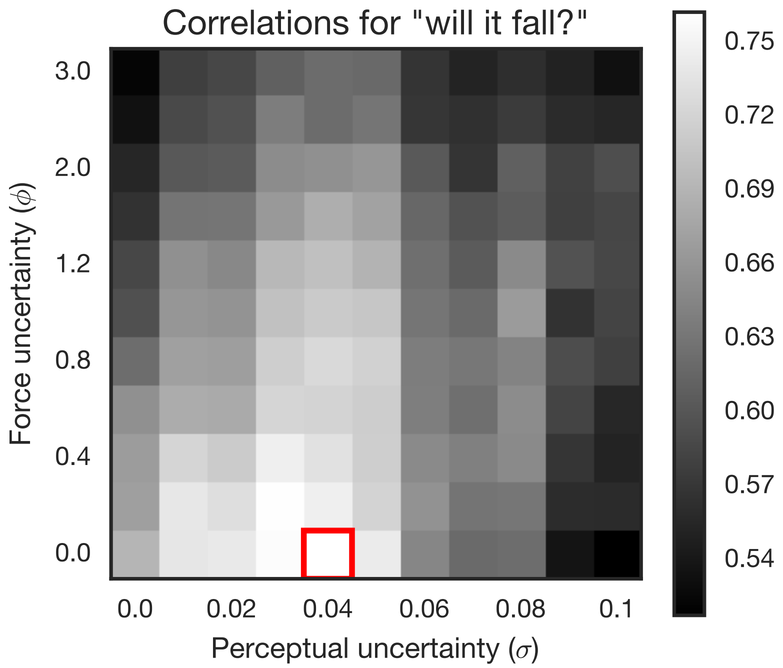


Figure A.1: Correlation as a function of model parameters. We queried the IPE model for ``will it fall?'' responses for different settings of its parameters, finding a similar pattern as that from Battaglia et al. (2013).

certainty, and ϕ represents force uncertainty. We then computed correlations between people's responses to "will it fall?" and the model's responses for each one of these parameter combinations (Figure A.1). The parameters that best fit people's responses to "will it fall?" were $\sigma = 0.04$ and $\phi = 0.0$. Battaglia et al. (2013) found $\sigma = 0.04$ and $\phi = 0.2$ to be the best fitting parameters; because these are close to our best fit parameters, we opted to perform all our analyses with $\sigma = 0.04$ and $\phi = 0.2$.

A.3 IPE Model Queries

What does it mean for a tower to "fall down"? One interpretation is that at least one block moved:

$$p(F_t|S_t, \kappa = k) \approx \frac{(\sum_{i=1}^{N} I_{b^{(i)}>0)}) + 0.5}{N+1}, \tag{A.5}$$

Table A.1: Log-likelihood ratios and Bayes factors for different query types.

		Exp. 4.1	Exp. 4.2	Exp. 4.3 within subjs.	Exp. 4.3 between subjs.
LLR	At least one	34.50	77.35	49.22	63.28
	More than half	-32.39	46.70	27.91	26.29
	Percent	-38.74	56.51	43.42	71.25
$\log K$	At least one	-163.41	9.51	0.56	26.72
	More than half	-247.21	-41.23	-25.44	11.89
	Percent	-217.81	0.04	-0.52	26.66

where $b^{(i)}$ is the proportion of blocks that moved during the i^{th} simulation, and where $I_{b^{(i)}>0}$ is one when at least one block has moved, and zero otherwise.

Another interpretation is that several blocks moved:

$$p(F_t|S_t,\kappa=k)\approx\frac{(\sum_{i=1}^{N}I_{b^{(i)}>0.5})+0.5}{N+1}, \tag{A.6}$$

where $I_{b^{(i)}>0.5}$ is one when more than half the blocks moved, and zero otherwise.

Yet another interpretation is that the tower falls "more" as the number of blocks that moved increases:

$$p(F_t|S_t,\kappa=k)\approx\frac{1}{N}\sum_{i=1}^{N}b^{(i)}, \tag{A.7}$$

where as before $b^{(i)}$ is the proportion of blocks that moved during the i^{th} simulation.

The "queries" presented in the previous paragraph are all plausible ways that people compute the answer to "will it fall?". Looking at correlations between people's responses to "will it fall?" in Experiments 4.1-4.2 and model responses, we found that the "at least one" query had a correlation of $r = 0.62$, 95% CI $[0.48, 0.75]$; the "more than half" query had a correlation of $r = 0.73$, 95% CI $[0.55, 0.85]$; and the "percent" query had a correlation of $r = 0.75$, 95% CI $[0.57, 0.87]$.

All of the queries also do a good job at predicting people's inferences in Experiment 4.1. The IPE observer model using the "at least one" query agreed with participants' inferences of which color was heavier 78.5% of the time. The correlation coefficient between the model's probability of choosing $\kappa = 10$ and the proportion of people that chose $\kappa = 10$ was $r = 0.63$, 95% CI $[0.47, 0.75]$. For the "more than half" query, the model agreed with people 81.9%

of the time, and the correlation coefficient was $r = 0.89$, 95% CI $[0.83, 0.93]$. For the "percent" query, the model agreed with people 81.9% of the time, and the correlation coefficient was $r = 0.91$, 95% CI $[0.87, 0.94]$.

If we fit the IPE learning and static observer models with each of the queries to participants' data, we also find similar log-likelihoods for each of the queries, with the exception of the "at least one" query for Experiment 4.1 (Table A.1). The Bayes factors for each query are also consistent, including the "at least one" query for Experiment 4.1. Based on these results, we suspect the anomalous log-likelihood ratio is due to overfitting, which the Bayes factors do not suffer from.

A.4 Derivation of the Counterfactual Likelihood

Here we derive Equation 4.9. First, we have two binary variables: the mass ratio (κ), and whether the tower falls or not (F). Based on how the towers were constructed, we know that if $F = 1$ and $\kappa = k$, then $F = 0$ and $\kappa \neq k$ (and vice versa; if $F = 0$ and $\kappa = k$, then $F = 1$ and $\kappa \neq k$). However, our simulation model only allows us to sample independently from the marginal probability distributions $p(F_{\kappa=k})$ and $p(F_{\kappa\neq k})$, where $F_{\kappa=k}$ means the feedback obtained when $\kappa = k$ and $F_{\kappa\neq k}$ means the feedback obtained when $\kappa \neq k$ Thus, our joint samples will follow the proposal distribution:

$$g(F_{\kappa=k}, F_{\kappa\neq k}) = p(F_t|S_t, \kappa = k)p(F_t|S_t, \kappa \neq k).$$

The acceptance probability is:

$$q(F_{\kappa=k}, F_{\kappa\neq k}) = \begin{cases} 1, & F_{\kappa=k} \neq F_{\kappa\neq k} \\ 0, & F_{\kappa=k} = F_{\kappa\neq k} \end{cases}$$

According to the definition of rejection sampling, the true value of the joint distribution $p(F_{\kappa=k}, F_{\kappa\neq k})$ is:

$$p(F_{\kappa=k}, F_{\kappa\neq k}) \propto q(F_{\kappa=k}, F_{\kappa\neq k})g(F_{\kappa=k}, F_{\kappa\neq k}).$$

As in the main text, letting $\mathcal{L}(\kappa) := p(F_t = 1|S_t, \kappa = k) = 1 - p(F_t = 0|S_t, \kappa \neq k)$, we can expand this out to get the full distribution:

$$
\begin{aligned}
p(F_{\kappa=k} = 0, F_{\kappa\neq k} = 0) &= 0, \\
p(F_{\kappa=k} = 1, F_{\kappa\neq k} = 1) &= 0, \\
p(F_{\kappa=k} = 1, F_{\kappa\neq k} = 0) &= \frac{1}{Z} \cdot \mathcal{L}(\kappa)\left(1 - \mathcal{L}(\bar{\kappa})\right), \\
p(F_{\kappa=k} = 0, F_{\kappa\neq k} = 1) &= \frac{1}{Z} \cdot \left(1 - \mathcal{L}(\kappa)\right)\mathcal{L}(\bar{\kappa}),
\end{aligned}
$$

where the normalization constant is:

$$
Z = \mathcal{L}(\kappa)\left(1 - \mathcal{L}(\bar{\kappa})\right) + \left(1 - \mathcal{L}(\kappa)\right)\mathcal{L}(\bar{\kappa}).
$$

A.5 Software

The tower stimuli were rendered using the Panda3D video game framework (Panda3D Developers, 2013), and were simulated using the Bullet physics engine (Bullet Developers, 2013). All analyses were performed in the Python programming language using the following scientific computing libraries: NumPy and SciPy (van der Walt, Colbert, & Varoquaux, 2011), Pandas (McKinney, 2010), and IPython (Pérez & Granger, 2007). Figures were generated with Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2014).

B

Chapter 5 Details

B.1 Experiment B.1: Replication of Smith & Vul (2013)

Because our experiments were performed online, we needed to fit the parameters of the model from Smith and Vul (2013) to reflect these different viewing conditions. To do this, we performed an online replication of the experiment from Smith and Vul (2013) in which we asked people to catch a ball like the one in Figure 6.1 using a paddle that could move up and down along the y-axis.

B.1.1 Participants

We recruited 60 participants using psiTurk (Gureckis et al., 2015). Participants were treated in accordance with UC Berkeley IRB standards and were paid $0.60 for 5 minutes of work. We excluded 18 people for failing to catch the ball on more than one control trial.

B.1.2 Stimuli

The stimuli were modified versions of those used in Experiments 6.1-6.3, with two differences. First, instead of a wall with a hole in it, there was a paddle of length 100px that could move up and down the y-axis. Second, instead of a full feedback animation, we just displayed the last

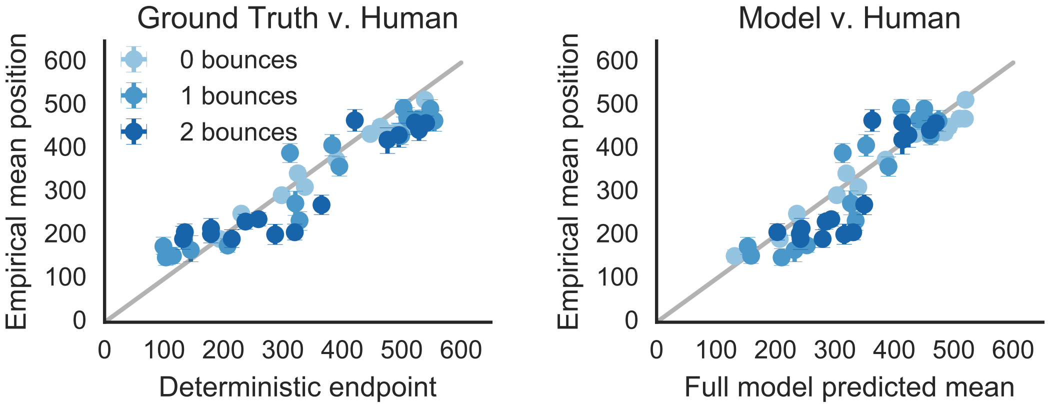


Figure B.1: Replication of Smith & Vul (2013). In both subplots, each point corresponds to a different stimulus, and colors correspond to the number of times the ball bounces under occlusion. The x-axis always corresponds to the model predicted endpoint, and the y-axis always corresponds to the average human predicted endpoint. *Left*: ground truth physics model versus human predictions. *Right*: noisy, approximate physics model versus human predictions.

frame. Because there was no hole that could vary by trial, there were only 48 stimuli, plus seven instruction and eight catch trials.

B.1.3 Procedure

Like Experiments 6.1-6.3, there were two phases: the training phase and the experimental phase. On each trial, participants were shown the scene, including the initial position of the ball. The paddle begin at the center of the y-axis, and was freely movable at the start of the trial. Participants were instructed to press "space" to begin the trial and display the stimulus animation. After the stimulus presentation, a gray occluder appeared, as well as a timer that began counting down for 2 seconds. During this time, participants had to move the paddle to catch the ball in the position it would be when the timer was up. When the timer finished, the paddle froze, the occluder was removed, and the full path of the ball was revealed. Participants were told whether they caught the ball or not, and then instructed to press "space" to begin the next trial.

B.1.4 Results

We fit the model parameters of σ_p, κ_v, κ_m, κ_b, and σ_0 to participant's responses (for details, see Smith & Vul, 2013), finding the best fitting parameters to be $\sigma_p = 0.003$, $\kappa_v = 94.03$, $\kappa_m = 2561041.94$, $\kappa_b = 46.66$, and $\sigma_0 = 155.24$. With these parameters, we found very similar results to those from Smith and Vul (2013), which are shown in Figure B.1. We found a correlation of $r = 0.95,\ 95\%$ CI $[0.92, 0.98]$ between people's average location of the paddle and ground-truth physics, which was slightly (though not significantly higher) than the correlation with the fitted model, which was $r = 0.92,\ 95\%$ CI $[0.88, 0.95]$. However, the fitted model resulted in a lower mean absolute error (MAE), which was $\mathrm{MAE} = 40.67,\ 95\%$ CI $[32.36, 49.78]$ compared to $\mathrm{MAE} = 46.63,\ 95\%$ CI $[38.97, 54.55]$ for the deterministic endpoint.

B.2 Experiment B.2: Determining Factors of Response Time

To begin to explore how people allocate their cognitive resources, we first wanted to establish the strategy by which people solve the types of physical reasoning problems such as the one in Experiment B.1. The simulation-based model from Smith and Vul (2013) gave a good fit to people's responses, but does not entirely rule out the possibility that people were relying on an alternate strategy. A stronger case could be made for the use of mental simulation if people's response times also vary as a result of the physical properties of the stimulus: for example, the mental simulation model would predict that people should take longer to respond when the ball has to travel a further distance. To test this hypothesis, we designed an experiment very similar to the original "paddle" experiment, except that rather than catching the ball, participants had to predict whether the ball would go through a hole or not. This change was made in order to remove the effect of differences in response times due to differences in action times, rather than thinking times.

B.2.1 Participants

We recruited 40 participants on Amazon's Mechanical Turk using the psiTurk (Gureckis et al., 2015) experimental framework. Participants were treated in accordance with UC Berkeley IRB standards and were paid $0.70 for 6.5 minutes of work. Participants were randomly assigned to one of eight conditions, which determined which stimuli they judged (see Appendix B.2.2). We excluded 1 participants for answering incorrectly on more than one control trial (see Stimuli), leaving a total of 39 participants.

B.2.2 Procedure and Stimuli

The procedure was identical to that of Experiments 6.1-6.3. The stimuli also had the same format, with a few exceptions. There were 48 different initial animations, equally balanced by number of bounces during feedback (24 each for 0 or 1 bounces). For each of these initial animations, there were two trial types and four path lengths, for a total of eight versions of each stimulus. The two trial types were: "far in" (FI), where the ball went through the center of the hole; and"far miss" (FM), where the ball missed the hole by a wide margin. The path lengths corresponded to the amount of time the ball traveled under occlusion, which was either 0.6s, 0.9s, 1.2s, or 1.5s. The hole locations and sizes were determined for each version of each stimulus, such that the probability of going in the hole was approximately either 0.9 (for "far in" trials) or 0.1 (for "far miss" trials).

In order to ensure that participants never saw the same initial animation twice, we used a Latin square design of Initial Animation $\times$ Trial Type $\times$ Path Length. Thus, each participant saw each initial animation once, each trial type 24 times, and each occlusion time 12 times. This also ensured that the ball would go through the hole half the time, so that participants would not be biased to respond either way. Additionally, there were seven instruction trials and eight control trials, which were the same for all participants and which were the same as those in Experiments 6.1-6.3. Thus, participants saw a total of 63 trials.

B.2.3 Results

B.2.3.1 Responses

On average, participants were correct $90.1\%,\ 95\%$ CI $[88.7\%, 91.4\%]$ of the time[1], responding that the ball would go in the hole on $93.2\%,\ 95\%$ CI $[91.5\%, 94.7\%]$ of "far in" trials, and $13.1\%,\ 95\%$ CI $[10.9\%, 15.2\%]$ of "far miss" trials. Thus, participants closely matched the desired 90% and 10% response rates we had intended when designing the stimuli. To determine what factors accounted for differences in accuracy, we constructed a mixed effects binomial GLM with fixed effects terms for the number of bounces, the trial type, the path length, and their interactions. We used the initial stimulus presentation and participant id as random effects terms. We found significant main effects for the number of bounces ($\chi^2(1) = 4.733, p < 0.05$) and the path length ($\chi^2(1) = 5.878, p < 0.05$), but not the trial type ($\chi^2(1) = 1.576, p = 0.21$).

[1] All averages are reported along with 95% bootstrapped confidence intervals computed using 10000 bootstrap samples (with replacement).

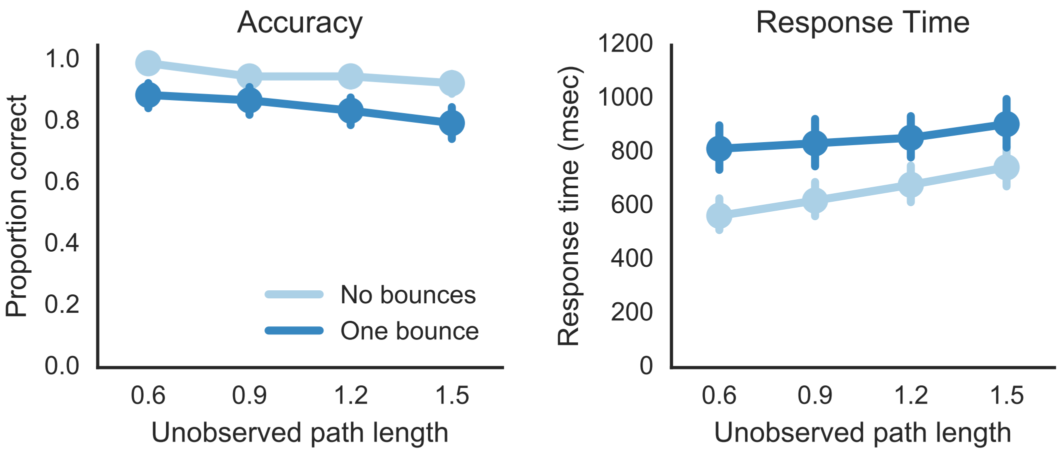


Figure B.2: Responses and RTs as a function of path length. Left: average accuracy as a function of unobserved path length and the number of bounces. Right: average response times as a function of unobserved path length and the number of bounces. In both cases, the path length and the number of bounces have a strong linear effect.

Additionally, we did not find any significant interaction effects. These results are shown in Figure B.2.

B.2.3.2 Response times

In all response time analyses, we computed means over log-transformed response times. Participants took $RT = 744.06\,\mathrm{msec},\ 95\%\ \mathrm{CI}\ [718.21, 771.09]$ to respond across all trials, with participants responding more quickly on "far in" trials ($RT = 681.82\,\mathrm{msec},\ 95\%\ \mathrm{CI}\ [648.83, 715.72]$) than on "far miss" trials ($RT = 812.30\,\mathrm{msec},\ 95\%\ \mathrm{CI}\ [771.93, 854.76]$). To determine what other factors accounted for differences in response times, we constructed a linear mixed effects model for log-transformed response times, with fixed effects terms for the number of bounces, the trial type, the path length, and their interactions. We used the initial stimulus presentation and participant id as random effects terms. We found significant main effects for path length ($\chi^2(1) = 17.344, p < 0.001$), the number of bounces ($\chi^2(1) = 37.495, p < 0.001$), and the trial type ($\chi^2(1) = 7.115, p < 0.01$), as well as interactions between the number of bounces and trial type ($\chi^2(1) = 12.780, p < 0.001$), between the number of bounces and path length ($\chi^2(1) = 11.210, p < 0.001$), and a three-way interaction between bounces, path length, and trial type ($\chi^2(1) = 5.875, p < 0.05$).

B.2.3.3 *Relationship of responses and RTs*

As noted in Appendix B.2.2, we tried to control for an effect of difficulty on response times by making all stimuli have a probability of around $p = 0.90$ or $p = 0.10$ of going in the hole. However, it is still possible that small variations in stimulus difficulty could have influenced people's response times by causing them to do more computation (i.e., run more simulations). To test for this, we constructed a linear mixed effects model for average log-transformed response times as a function of path length, trial type, and the number of bounces (as well as their interactions). We compared this model to one that additionally included a term for people's average responses. We did not detect a difference between these two models ($\chi^2(1) = 1.860, p = 0.17$), indicating that there was no discernible effect of difficulty on people's response times.

B.2.3.4 *Learning*

To check for practice effects, we computed Spearman rank correlations (with 95% confidence intervals computed from 10000 bootstrap samples) between trial number and accuracy, as well as between trial number and log-transformed response times. We found a small overall effect of practice on accuracy ($\rho = 0.34,\ 95\%\ \mathrm{CI}\ [0.02, 0.42]$), though individually this was not significant in either the first ($\rho = 0.29,\ 95\%\ \mathrm{CI}\ [-0.09, 0.48]$) or second ($\rho = 0.00,\ 95\%\ \mathrm{CI}\ [-0.25, 0.28]$) halves of the experiment. There was also an overall effect of practice on RT ($\rho = -0.56,\ 95\%\ \mathrm{CI}\ [-0.55, -0.25]$), though this effect disappeared in the second half of the experiment ($\rho = -0.35,\ 95\%\ \mathrm{CI}\ [-0.40, 0.00]$).

B.2.4 Discussion

The results from Experiment 6.2 reveal that when predicting whether the ball will go into the hole, people's responses and response times are strongly influenced by both the number of times the ball will bounce as well as well as the path length of the ball under occlusion. Because both of these are *unobserved* properties of the ball's trajectory, these results give weight to the hypothesis that people are indeed running mental simulations to solve the task. One might argue that the path length has an effect on response times due to differences in saccade distances, rather than mental simulation per se. However, the fact that the number of bounces has such a strong effect argues otherwise: if people were purely relying on surface features of the visual stimulus, then *a priori* there is no reason why the number of bounces should affect their response times.

B.3 SPRT Proofs

B.3.1 Derivation of Equation 6.4

From Feller (1968, Eq. 2.4, p. 345), we have:

$$p(r = 0 \mid q, p, a, z) = \frac{(q/p)^a - (q/p)^z}{(q/p)^a - 1}, \tag{B.1}$$

where $q = 1 - p$, z is the starting value of the accumulator, and the absorbing boundaries are at 0 and a. In our case, we have $z = T$ and $a = 2T$. Thus, we can rewrite Equation B.1 as:

$$p(r = 1 \mid T, p) = \frac{(q/p)^T - 1}{(q/p)^{2T} - 1}.$$

Now, expanding out the denominators of the fractions:

$$p(r = 1 \mid T, p) = \frac{\frac{q^T - p^T}{p^T}}{\frac{q^{2T} - p^{2T}}{p^{2T}}} = \frac{(q^T - p^T)p^T}{q^{2T} - p^{2T}} = \frac{(q^T - p^T)p^T}{(q^T + p^T)(q^T - p^T)} = \frac{p^T}{(1-p)^T + p^T}.$$

B.3.2 Derivation of Equation 6.5

From Feller (1968, Eq. 5.7, p. 353), we have:

$$p(N, r = 0 \mid q, p, a, z) = \\ a^{-1} 2^N p^{(N-z)/2} q^{(N+z)/2} \sum_{v=1}^{a-1} \cos^{N-1}\left(\frac{\pi v}{a}\right) \sin\left(\frac{\pi v}{a}\right) \sin\left(\frac{\pi z v}{a}\right), \tag{B.2}$$

where $q = 1 - p$, z is the starting value of the accumulator, and the absorbing boundaries are at 0 and a. In our case, we have $z = T$ and $a = 2T$, allowing us to write:

$$C_{N,T} = \frac{2^N}{2T} \sum_{v=1}^{2T-1} \cos^{N-1}\left(\frac{\pi v}{2T}\right) \sin\left(\frac{\pi v}{2T}\right) \sin\left(\frac{\pi v}{2}\right),$$

$$p(N, r = 0 \mid T, p) = C_{N,T} \cdot p^{\frac{N-T}{2}} (1-p)^{\frac{N+T}{2}}.$$

Because the absorbing boundaries are symmetric, we have $p(N, r = 1 \mid T, p) = p(N, r = 0 \mid T, 1 - p)$. Using this identity, we can marginalize over values of r, obtaining:

$$p(N \mid T, p) = C_{N,T} \cdot \left[p^{\frac{N-T}{2}} (1-p)^{\frac{N+T}{2}} + (1-p)^{\frac{N-T}{2}} p^{\frac{N+T}{2}} \right].$$