Title

Principles underlying human physical prediction

Author

Smith, Kevin

Publication Date

2015

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Principles Underlying Human Physical Prediction

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Psychology

by

Kevin A. Smith

Committee in charge:

Professor Edward Vul, Chair
Professor David Barner
Professor Benjamin Bergen
Professor Roger Levy
Professor Donald MacLeod

2015

The dissertation of Kevin A. Smith is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

TABLE OF CONTENTS

Signature Page iii

Table of Contents iv

List of Figures vi

List of Tables viii

Acknowledgements ix

Vita xi

Abstract of the Dissertation xiii

1 Introduction 1
1.1 Prior research on ‘intuitive physics’ 3
1.2 Simulation-based physical reasoning 6
1.3 Defining the structure of physical simulation 10
1.4 Conclusion 14
References 15

2 Radically different physical intuitions in action and conception 20
2.1 Introduction 21
2.2 Experiment 1: Difference between interaction and conception 24
2.3 Experiment 2: The impact of stimulus richness on physical knowledge . 38
2.4 Discussion 45
References 47

3 Sources of uncertainty in intuitive physics 50
3.1 Introduction 50
3.2 Stochastic physics model 52
3.3 Experiment 54
3.4 Model application 58
3.5 Discussion 69
References 71

4 Accumulating physical evidence over time 72
4.1 Introduction 73
4.2 Experiment 75
4.3 Physical evidence accumulation 79
4.4 Alternative decision heuristics 84
4.5 Discussion 90

References 93

5 General discussion 95

5.1 Modes of physical reasoning 97

5.2 Approximations of physics 104

5.3 Conclusion 115

References 115

A Supplemental materials for Chapter 2 119

A.1 Model details 119

A.2 Supplementary results 123

LIST OF FIGURES

Figure 2.1: Diagram of trials in the catching (*top*) and releasing (*bottom*) trials. 26

Figure 2.2: Handout provided to participants for the drawing task. Instructions and stimuli were based on Caramazza et al. (1981). 28

Figure 2.3: Diagrams of the forward model predictions for the path of the ball at four different cut points along the pendulum, using best fitting parameters. 32

Figure 2.4: Illustration of the noisy Newton account of human judgments in the catching and releasing tasks. 33

Figure 2.5: *Left*: The bias and variance of participants' average performance on the interactive (catching and releasing) tasks is better captured by the noisy Newton model than ground truth or the angled non-physical model. *Right*: Classification of responses on the drawing task. . . . 35

Figure 2.6: Fits of each model to human catching and releasing behavior. . . . 37

Figure 2.7: *Top*: Diagram of a trial. *Bottom*: The six potential paths raters could classify each drawing as (not including *unclassified*). 41

Figure 3.1: Sources of uncertainty in the stochastic physics model. 53

Figure 3.2: Diagram of a trial. 55

Figure 3.3: Mean predicted paddle position versus path endpoint using deterministic physics as a function of trial condition. Each point represents a separate trial. 57

Figure 3.4: Sample simulation paths for one trial with each model. 62

Figure 3.5: Full model predictions vs. empirical mean position by condition. . . 63

Figure 3.6: Full model vs. empirical standard deviation by trial. Each point represents a separate trial. 65

Figure 4.1: *Left:* Screenshot of a trial. *Right:* The path that the ball traveled on the example trial. 77

Figure 4.2: Joint histogram of model (x-axis) and human (y-axis) decisions. . . 82

Figure 4.3: Participants' average scores on a trial versus predicted scores from the physical simulation model. 83

Figure 4.4: Sample trials with changes in confidence. 85

Figure 4.5: Joint histogram of model (x-axis) and human (y-axis) decisions for all models. 89

Figure 4.6: Contained trials. 92

Figure A.1: Measures of the coefficient of determination for each participant in both the catching and releasing tasks. 124

Figure A.2: Histogram of error correlations between each subject and all other subjects across all three tasks in Experiment 1. 125

Figure A.3: Plots of correlations between each participant's extrapolated drawings and the average of all other subjects. 126
Figure A.4: Individual model comparisons. 128

LIST OF TABLES

Table 2.1: Individual best fitting model vs. classification on drawing task. . . . 37

Table 2.2: Classification of participants drawings, split by pendulum cut point and experimental condition. 43

Table 3.1: Percent of distance ‘shifted’ from actual end ball position towards center by trial condition 58

Table 3.2: Average standard deviation (in pixels) of responses within a trial by condition 58

Table 3.3: Model prediction of left-out data 62

Table 3.4: Correlation between model and empirical by-trial means within condition 64

Table 3.5: Correlation between full model and empirical by-trial standard deviations within condition 66

Table 3.6: Autocorrelation of paddle position with prior position 67

Table 4.1: Predictors used in the surface-level feature model 87

ACKNOWLEDGEMENTS

First and foremost, I'd like to thank my advisor Edward Vul for his excellent guidance throughout my graduate career. It is only due to his advising that I have learned how to ask interesting scientific questions, think critically about my own research, and navigate the corridors of academia. I could not have asked for a better mentor.

Thank you also to the other members of my committee – David Barner, Benjamin Bergen, Roger Levy, and Donald MacLeod – with whom I have had many interesting discussions about both this work in particular, and psychological sciences in general. I am also grateful to all of my coauthors, not just those that have contributed directly to this thesis – Joshua Tenenbaum, Peter Battaglia, and Eyal Dechter – but everyone who has contributed to the thoughts and ideas that have formed the framework for my research.

The other graduate students and post-docs in my lab, past and present, have also been instrumental in my graduate career. Thank you to Drew Walker, Timothy Lew, Kristin Donnelly, Robert St. Louis, Nisheeth Srivastava, and Cory Reith for providing feedback, discussing ideas, or simply helping me unwind after a long day of research. And thank you to Banjo, who always took me on walks to get coffee when my energy was flagging.

Of course, I wouldn't be where I am without the support of my family and friends. So thank you to my parents who have not just supported me in all my endeavors but also taught me from an early age how to make clear, logical arguments, and thank you to my brothers – Dylan, Eric, and Ross – who have provided good sparring partners for those arguments. And a special thanks to Melissa Troyer, who has had to deal with my neuroses as I have finished this thesis, and yet has still supported me the entire way.

Finally, because it is a university requirement, I'd like to acknowledge the various publications that have gone into this thesis:

Chapter 2, in full, is currently under review for publication of the material in

Cognition. Smith, Kevin A; Battaglia, Peter W; Vul, Edward. The thesis author was the primary investigator and author of this material.

Chapter 3, in full, is a reprint of the material as it appears in *Topics in Cognitive Science* 5(1), 2013. Smith, Kevin A; Vul, Edward. The thesis author was the primary investigator and author of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Smith, Kevin A; Dechter, Eyal; Tenenbaum, Joshua B; Vul, Edward. The thesis author was the primary investigator and author of this material.

VITA

2005	B. A. in Cognitive Science *magna cum laude*, Dartmouth College
2011	M. A. in Psychology, University of California, San Diego
2010-2015	Graduate Teaching Assistant, University of California, San Diego
2015	Ph. D. in Psychology, University of California, San Diego

PUBLICATIONS

Kevin A Smith, Edward Vul "The role of sequential dependence in creative semantic search", *Topics in Cognitive Science*, 7(3), 2015.

Kevin A Smith, Edward Vul "Prospective uncertainty: The range of possible futures in physical predictions", *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015.

Jessica B Hamrick, Kevin A Smith, Thomas L Griffiths, Edward Vul "Think again? The amount of mental simulation tracks uncertainty in the outcome", *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015.

Drew E Walker, Kevin A Smith, Edward Vul "The 'Fundamental Attribution Error' is rational in an uncertain world", *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015.

Kevin A Smith, Edward Vul "Reductionism and practicality", *Cosmos and History: The Journal of Natural and Social Philosophy*, 10(1), 2014.

Kevin A Smith, Edward Vul "Looking forwards and backwards: Similarities and differences in prediction and retrodiction", *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2014.

David D Bourgin, Joshua T Abbott, Thomas L Griffiths, Kevin A Smith, Edward Vul "Empirical evidence for Markov Chain Monte Carlo in memory search", *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2014.

Kevin A Smith, David E Huber, Edward Vul "Multiply-constrained semantic search in the Remote Associates Test", *Cognition*, 128(1), 2013.

Kevin A Smith, Peter W Battaglia, Edward Vul "Consistent physics underlying ballistic motion prediction", *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 2013.

Kevin A Smith, Eyal Dechter, Joshua B Tenenbaum, Edward Vul “Physical predictions over time”, *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 2013.

Kevin A Smith, Edward Vul, “Sources of uncertainty in intuitive physics”, *Topics in Cognitive Science*, 15(1), 2013.

Kevin A Smith, Edward Vul, “Sources of uncertainty in intuitive physics”, *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 2012.

ABSTRACT OF THE DISSERTATION

Principles Underlying Human Physical Prediction

by

Kevin A. Smith

Doctor of Philosophy in Psychology

University of California, San Diego, 2015

Professor Edward Vul, Chair

Our days are filled with instances of reasoning about the physics of the world, from simple tasks such as stacking dishes in a way that keeps them stable, to life-and-death decisions such as not crossing the street because we presume an oncoming car would hit us if we did. Yet the process we use to make inferences about physical events is not well understood. Here I argue that these interactions are based on a rich, approximately accurate simulation of physical events, but we must account for uncertainty about the current properties of objects in the world. In this thesis I investigate the structure of this simulation process and how it relates to other facets of cognition, including (1) demonstrating that the principles underlying interactions with the world are based on accurate physics, even if our explanations of those same principles are idiosyncratic and erroneous, (2) mapping out the types of uncertainty that this process accounts for, and demonstrating that the simulations themselves are therefore stochastic, and (3) explaining how physical predictions are updated over time due to changing evidence from evolving simulations. This provides a framework for understanding how people form and update representations of both the current and future state of the world based on rich, structured, probabilistic reasoning.

1 Introduction

From simple tasks like knowing where a wrapper will land when we throw it towards the trash can, to matters of life and death such as choosing not to cross the street when we might be hit by a car, our day-to-day actions require predicting what will happen next. Our continued survival is testament to the fact that the predictions we make about how the world will unfold are relatively accurate.

One crucial component of our predictions is reasoning about physical events – how objects will move about and interact over time. Physical reasoning underlies a wide range of our abilities that allow us to interact with the world, from knowing how much force to use to throw a ball to our friends, to judging whether a coffee cup perched on the edge of a table will fall, to pouring tea from a full kettle into a cup without spilling. It also supports a large range of other core human abilities – such as how we infer other peoples' goals by accounting for the physical constraints they are subject to when they plan their actions (Gergely & Csibra, 2003), or how we identify objects based on their physical relationships within a scene (Biederman, Mezzanotte, & Rabinowitz, 1982). Yet despite the importance of physical reasoning in our lives, how we perform these tasks is often mischaracterized or not well understood.

Prior research over the past 35 years has often characterized our understanding and use of physical principles as deficient. Some theories claim that our physical reasoning is based on a pre-Newtonian 'impetus theory' (e.g., McCloskey, 1983a), while other theories propose that we can only reason about one facet of physics at a time (e.g., Proffitt & Gilden, 1989), and yet others suggest that we can only account for qualitative relationships between objects (e.g., Forbus, 1994). Due to this extensive research, an underlying assumption about human physical reasoning is that it is 'good enough' for us to function, but that our understanding of physics does not reflect the way that physics works in reality.

More recently, however, evidence has been accumulating that suggests that human physical reasoning is approximately correct, and that errors in our physical predictions are a result of accounting for uncertainty about the state of the world – including object locations, motions, and properties (Sanborn, Mansinghka, & Griffiths, 2013). According to this hypothesis, termed the 'noisy Newton' theory of physical reasoning, people use an 'approximate physics engine' to run the world forward and propose hypothesized physical outcomes (Battaglia, Hamrick, & Tenenbaum, 2013). While the broad framework of this theory has been proposed, there are many outstanding questions about how this physics engine functions: what physical rules does it use in the course of its simulations? What sorts of uncertainty must it account for? How do we integrate information from simulations when the world itself is changing?

In this thesis I will first reconcile the noisy Newton theory with the prior literature on intuitive physics, and then describe how people use physical simulation to inform their predictions about the future state of the world. In this chapter I provide a brief history of theories about intuitive physics, and how they lead to the questions answered in this thesis. In Chapter 2 I show that the dichotomy between classical and recent studies of intuitive physics can be explained by task demands: when people interact with the world their predictions are based on relatively accurate principles, but when they reason explicitly about how the world should work, their conceptions are idiosyncratic and often incorrect in the ways described in prior literature. In Chapter 3 I show that people incorporate not just uncertainty about the location and motion of objects in their physical predictions, but also consider uncertainty about more nuanced properties of objects – e.g., that a ball may not bounce cleanly off a wall because of imperfections in either the ball or the wall. In Chapter 4 I demonstrate that the noisy Newton framework can explain not just how people make singular predictions, but how those predictions evolve in real-time by accumulating changing evidence from an approximate physics engine that is updated as the world itself changes. Finally, in Chapter 5 I discuss how these experiments form a framework for understanding physical reasoning, and possible future directions for expanding this framework.

1.1 Prior research on 'intuitive physics'

1.1.1 Psychological theories of intuitive physics

One of the earliest studies into the structure of physical reasoning found what was at the time a very curious result: while a ball that exits a curved tube should travel in a straight line (there is no longer any force from the tube to cause it to curve), a proportion of people who are asked about this scenario believe that the ball will continue to move on a curved path because it retains some of the curving motion from the tube (McCloskey, Caramazza, & Green, 1980; McCloskey & Kohl, 1983). Continued research found a number of additional errors that people make: for instance, we cannot accurately describe the ballistic trajectory of an object (Caramazza, McCloskey, & Green, 1981; Hecht & Bertamini, 2000), we incorrectly believe that an object dropped from a moving carrier (e.g., dropping a pen while walking) will lose all forward velocity and fall straight down (McCloskey, Washburn, & Felch, 1983), and we inaccurately claim that the water line will tilt when we pour liquid from a glass (Howard, 1978; Hecht & Proffitt, 1995).

There have been a number of theories proposed to explain these errors in physical reasoning. The most well known of these theories, ‘impetus theory,’ suggests that human physical knowledge is inaccurate, based on pre-Newtonian theories that involve concepts such as impetus – the incorrect notion that forces must continuously be applied to keep an object in motion (McCloskey, 1983a, 1983b; Kozhevnikov & Hegarty, 2001). While this theory characterized the modal responses on many of the tasks used to elicit these errors and generally followed the explanations that participants provided to support their responses, it is not clear that impetus theory is a universal human default for understanding physics. Even between scenarios that require reasoning about the same physical principles, peoples’ explanations are inconsistent – people will respond in a way that indicates impetus physics in one and accurate, Newtonian physics in another (e.g., determining the motion of a ball after it exits two curved tubes that differ only in terms of the amount of curvature; problems 1 & 2 from McCloskey et al., 1980); this may be because people can reason correctly about familiar instances, but use incorrect theories when presented with unfamiliar or abstract problems (Catrambone, Jones, Jonides, & Seifert, 1995; Kaiser, Jonides, & Alexander, 1986).

Others have proposed alternate theories to explain these errors in reasoning: that we can only attend to a single physical dimension at a time (Proffitt & Gilden, 1989), that we misrepresent the way that forces are imparted from one object to another (White, 2012), or that physical knowledge is combined ad-hoc from a collection of primitive notions about object interactions (diSessa, 1993; Halloun & Hestenes, 1985). However, all of these theories have a common thread: that human physical reasoning is at its core based on erroneous principles or otherwise flawed.

1.1.2 Qualitative physics

At the same time that psychologists were discovering errors in human physical reasoning, artificial intelligence researchers were attempting to build expert systems that could reason about physical events. This led to the development of ‘qualitative physics’ systems – systems that extract qualitative rather than quantitative information from a scene (e.g., that a car is slowing down rather than that a car is decelerating at $5.2m/s^2$), then apply rules based on those relations (Forbus, 1984; Gardin & Meltzer, 1989). These qualitative descriptions represent a scene using information similar to peoples’ verbal descriptions: the book is on the table, or the ball is in the box (Forbus, 1983). Because these relationships are qualitative, they allow for easy abstraction – for instance, all objects resting on tables will continue to remain stationary in the absence of forces.

Forbus and Gentner (1986) argue that this forms the foundation of intuitive human conceptions about physics: we abstract specific rules (e.g., “books stay on tables”) from our scene perception, then generate causal structures based on these facts (e.g., “books would fall, but tables prevent them from falling to the floor”). Finally, we use analogical reasoning to abstract across causes (e.g., “if object A is resting on object B, object B is preventing object A from falling”). Note that this is a naïve conception of physics: it tells us that books have a natural tendency to fall that the table is preventing, not that the table is pushing back with a force equal and opposite to gravity according to Newton’s third law of motion. Nonetheless, this explanation is similar to human judgments – naïve subjects typically do not claim that the table is exerting any force on the book (Brown, 1994). Therefore, similar to the theories used to explain the errors in physical judgments, the theory that peoples’ physical reasoning is based on qualitative principles suggests that the rules that underly this reasoning are at best coarse approximations of accurate, Newtonian physics.

However, qualitative physics systems have had mixed results in describing human physical reasoning. In some domains qualitative reasoning can explain how people conceptualize physical principles: for instance, answering descriptive questions about simple scenes (Friedman & Forbus, 2009) or reasoning about containment (Davis, Marcus,

& Chen, 2013). But explaining kinematics though qualitative reasoning has proven to be problematic, since without quantitative measures of velocity and position there are not clear ways to differentiate, for instance, whether a ball rolling up a hill will have enough speed to go over the crest, or whether it will roll back down (Forbus, 1980). This has led some to claim that the only way to represent kinematic relationships is to incorporate *some* quantitative measures of the state of objects (Forbus, Nielsen, & Faltings, 1987).[1] Thus, while qualitative reasoning may explain some aspects of human physical reasoning, it is unlikely to be the only way that people conceptualize physics.

1.2 Simulation-based physical reasoning

For nearly a half century, many researchers have studied the process of 'simulation' – constructing a mental representation of the world, then iteratively updating that representation over time to understand how the world will evolve, even under counterfactual assumptions (Kahneman & Tversky, 1981). This mental simulation process has been hypothesized to underlie a wide range of human capabilities, including causal reasoning (Wells & Gavanski, 1989), theory of mind (Gallese & Goldman, 1998), and language comprehension (Zwaan, 2003; Bergen & Wheeler, 2010). But importantly for this thesis, it has also been seen as the basis of many different types of physical reasoning.

1.2.1 Origins of physical simulation theory

Modern research into mental simulation began with a study which found that when people determine whether two objects are the same or different shapes, their reaction times grow linearly with the angle that one object would need to be rotated through to

[1] There have been attempts to represent kinematics as changes through qualitatively segmented space (c.f., Cohn & Hazarika, 2001). However, this research typically focuses on the spatial representations and often does not directly address the difficulties of kinematic descriptions.

match the other (Shepard & Metzler, 1971), suggesting that people are mentally rotating the shapes in the same way that they would physically rotate them to ensure they are the same when aligned. Later work studied how this simulation process supported more complex physical reasoning – for instance, how visuospatial abilities support the ability to simulate pulley systems (Hegarty & Sims, 1994) or how people induce rules about gear systems from their own mental simulations (Schwartz & Black, 1996).

While this set of research suggests that a proportion of physical reasoning is based on dynamic simulations of visual imagery, it does not describe the rules of the simulations that gave rise to this reasoning. Indeed, the research into the core rules of simulation is contentious, with some claiming that they are based on incorrect principles, while others find contrary evidence that simulation makes our physical judgments more accurate. For instance, Kozhevnikov and Hegarty (2001) claim that the rules of simulation are based on the same erroneous principles of 'impetus physics.' But this claim is based on evidence from the 'representational momentum' literature that memory for the location of objects is shifted based on the dynamics of impetus physics rather than Newtonian physics (e.g., larger objects will be displaced downwards more than smaller objects because gravity affects them more; Hubbard, 1997), and it is unclear whether these representational momentum phenomena arise from the dynamics of objects rather than perceptual biases (Kerzel, 2002). On the other hand, there is evidence that activating mental imagery by asking people to view or imagine a scene in motion makes them rely on more accurate physical principles than they would use for simple explanation (Frick, Huber, Reips, & Krist, 2005; Kaiser, Proffitt, Whelan, & Hecht, 1992). Therefore it remains an outstanding question of what rules are used to drive our mental simulations of physical processes.

1.2.2 The ‘noisy Newton’ theory

The most recent theory of physical reasoning starts from an alternate hypothesis – not that human knowledge of physics is intrinsically flawed, but rather that it is a close approximation of Newtonian mechanics. According to this theory, biases in physical reasoning are instead the result of reasoning under uncertainty using prior information about object properties. Hence, peoples’ reasoning is both noisy and based on Newtonian mechanics – the ‘noisy Newton’ hypothesis.

The initial success of this theory was tying together two facets of peoples’ judgments about colliding objects. First, when people observe two objects colliding and are asked which of the two is heavier, they are biased by information that is irrelevant to the task (the elasticity of the objects), leading many to argue that these judgments are based on limited information and heuristics (Todd & Warren, 1982; Gilden & Proffitt, 1989). Second, when one object collides with another, the first object stops, and the second moves, the first object is seen to have ‘caused’ the second object to move only if the subsequent velocity falls within a limited range (Michotte, 1963). Though these two judgments had previously been considered and explained separately, Sanborn et al. (2013) explained peoples’ judgments across both domains using a single model of physical reasoning. This model assumed that people used accurate, Newtonian reasoning about collisions (in contrast to, e.g., Todd & Warren, 1982), though predicted that errors and biases would arise as a result of initial uncertainty in both perception and knowledge about the objects’ masses and elasticity.

However, it would be untenable to assume that people reason about physical events by using equations from Newtonian mechanics to calculate the future positions and velocities of objects. Newtonian equations work for problems involving two objects, but despite centuries of work, physicists and mathematicians have not determined an analytical method for calculating the future state of a system with three or more objects if those objects are allowed to collide with one another, and many consider this to be impossible (Diacu, 1996).[2] Instead, most modern computer physics engines ‘simulate’ physics in an iterative fashion, updating the state of the world and calculating the effect of object interactions in brief time-steps (Millington, 2010). Crucially, these physics engines are not exact, but instead require some approximation of how objects should move between these time-steps.

This problem, combined with the prior simulation literature, led Battaglia et al. (2013) to propose that people use an ‘intuitive physics engine’ to perform physical reasoning in a fashion similar to a computer physics engine – determining how the future will unfold by incrementally updating the world according to approximately accurate physical principles. However, these simulations start with uncertainty about the current state of the world, and thus the outcome of the simulations is necessarily probabilistic, providing a range of potential futures that might occur. Battaglia et al. (2013) used this framework to explain how people make judgments about a physical system that involves a large number of objects and massively unconstrained types of collisions: whether and, if so, how a tower of multiple, balanced blocks will fall.

This framework has since been extended to capture a range of human judgments, e.g., how people integrate physical and perceptual information to judge the positioning of an object (Scarfe & Glennerster, 2014), how people make causal judgments about objects running into each other (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012), or how people predict how fluids will splash around a container (Bates, Yildirim, Tenenbaum, & Battaglia, 2015).

This theory at its core makes two strong claims about the process we use to extrapolate the world forward: that our predictions of the future are created in an iterative

[2] Even analytical solutions of systems without collisions are wildly impractical due to the potentially chaotic nature of these systems, requiring sums of millions of terms of a series for even a modicum of accuracy over short time spans (Diacu, 1996).

manner through simulation, and that the rules we use to iterate these simulations will provide a good approximation of the future state of the world. The noisy Newton theory therefore rules out certain sets of prediction processes.

For one, this precludes any cognitive process that can directly predict the state of the world at an arbitrary future time without determining intermediate states. While mathematicians have yet to find an algorithm that can do this at all (much less in a reasonable time), it is entirely possible that the mind can approximately calculate the future without simulation. However, without proof that this computation is in fact possible, it is unclear how such a system might work.

Similarly, the noisy Newton theory claims that our rules for updating our simulations approximate Newtonian mechanics. This claim is more contentious, with many studies finding that that our understanding of physical rules is erroneous (e.g., McCloskey, 1983a; Kozhevnikov & Hegarty, 2001). It is therefore an open question of why these studies find errors in physical reasoning if our physical simulation is at its core relatively accurate.

However, even within the structure of the noisy Newton theory, there are a broad class of processes that would satisfy these criteria (for instance, any reasonable computer physics engine), and only a narrow subset of these can describe human cognition. Defining the exact simulation process that underlies human physical reasoning will therefore require further research.

1.3 Defining the structure of physical simulation

This thesis aims to fill the largest gaps in our understanding of how we perform physical reasoning. First, we must understand why prior studies have found that sometimes our physical reasoning is accurate and calibrated, and other times it is idiosyncratic and erroneous – when do we use noisy Newton simulations versus other methods of reasoning? Next, we must define what is ‘noisy’ about the noisy Newton framework. Prior research has suggested that our predictions are variable because of initial uncertainty about the properties of objects, but has remained agnostic about whether the physical dynamics we use to extrapolate motion are deterministic, or whether they are themselves noisy and uncertain. Finally, we must understand how people use physical simulation to make decisions about the world. The information gained from simulations will necessarily fluctuate over time as the world provides new information, and therefore it is important to understand how we aggregate information from simulations to provide evidence for what might occur, and how that affects the timing of our decisions and evolution of our beliefs over time.

1.3.1 When do we use accurate simulations? (Chapter 2)

The research into intuitive physics has two sets of disparate findings. On the one hand research like Battaglia et al. (2013) finds that our physical judgments are based on accurate principles, but on the other hand studies such as Caramazza et al. (1981) find that our physical principles are clearly erroneous. One theory suggests that this is due to differences in the types of physical principles studied: Battaglia et al. (2013) asked people to judge stability and we have an accurate representation of that principle, while Caramazza et al. (1981) asked people to make judgements about ballistic motion – a concept that we do not have an accurate representation for (Marcus & Davis, 2013). If this is true, then we cannot have an approximately accurate simulation engine driving our physical predictions, but instead this suggests we rely on non-physical heuristics that represent correct physical principles for only a narrow set of problems.

However, there is an alternate possibility: that on some tasks we use physical simulation that is based on accurate principles, while on other tasks we use other cognitive systems for reasoning about physics. If this theory holds, then we can still have an approximate physics engine, but we also have other methods for conceptualizing physical principles. A core question of physical simulation is therefore which of these two theories is correct, and if we do only sometimes use accurate simulation, when do we do so?

To test these theories, I turn to a task that has classically been demonstrated to elicit errors in physical judgment: gauging the ballistic trajectory of a ball that has been cut from a swinging pendulum (Caramazza et al., 1981). The errors that people make in this task – both when drawing the predicted trajectory and in their explanations – have been used as evidence that people hold ‘impetus physics’ beliefs (McCloskey, 1983a).

I show, however, that these errors are dependent on how people use these principles. When people interact with the pendulum system (by catching the ball in a bucket or choosing to cut the pendulum string so that the ball hits a target), their predictions about the trajectory of the ball can be explained by accurate physical mechanics and sensory uncertainty, despite the fact that their responses to an identical task as in Caramazza et al. (1981) display a range of physical errors. This suggests that the physical model we use to determine our interactions with the world is based on approximately accurate physics, but when we perform explicit reasoning about physical principles our theories can be idiosyncratic and erroneous.

1.3.2 What is ‘noisy’ in the noisy Newton framework? (Chapter 3)

A central claim of the noisy Newton framework is that our predictions are influenced by uncertainty about the world. Prior studies have suggested that our predictions are noisy due to uncertainty about the initial properties of objects (such as location and motion), but have made the simplifying assumption that the physical simulation itself was deterministic and noiseless (Battaglia et al., 2013; Sanborn et al., 2013). However, there are a priori unknowable features of the world that can influence the actual paths of objects: a gust of wind might push a thrown ball onto a different course, or a ball rolling across a grass field might bounce upwards if it hits an unseen bump. If our simulations are deterministic, then we may have systematic biases in our predictions due to the inability to account for these unknowable events. To fully account for the variability in real-world physical motion, we must use noisy dynamics to produce the same variability in our simulations. It is therefore an empirical question of whether our simulations are stochastic to account for this uncertainty – is our physics itself noisy, or just our initial understanding of the world?

In this experiment, participants predicted the final location of a ball as it bounced around a computerized table under a variety of conditions, and I found that peoples' predictions can only be explained if their simulations include noise that accumulates throughout extrapolation. This suggests that the 'noise' in the noisy Newton framework comes both from initial uncertainty and from stochastic noise added to our simulations.

1.3.3 How do we integrate information from simulation over time? (Chapter 4)

In a changing environment, our physical predictions should shift over time to account for these changes. It is therefore important to understand how simulation supports these evolving inferences: do we gradually update our predictions by integrating old and new simulations, or do we refresh our predictions with an entirely new set of simulations? Integrating evidence from simulations suggests that we can be biased by our past predictions, but this process can also serve to stabilize our predictions over time to ensure we are not biased by moment-to-moment changes in simulation.

In order to understand how predictions evolve over time, we must study the process and timing of accumulating evidence from our simulations, which allows us to predict not just which decision people will make when performing physical reasoning, but how fast they will make it and how their belief changes over time. In one study of this process, Hamrick, Smith, Griffiths, and Vul (2015) found that in order to make a simple dichotomous physical decision ("will this ball go through a hole in the wall or bounce off if it continues its trajectory") people produce a limited number of simulations until they reach a certain level of net evidence in favor of one of the choices, and that this process could explain the variability in peoples' reaction times across a variety of conditions. But this study looked at only a single decision, and therefore could not explain how beliefs might shift in light of new evidence.

In the final experiment, I asked participants to watch a ball bouncing around a computerized table and continuously predict which of two 'targets' the ball would reach first – thereby getting a continuous measure of peoples decisions as they obtained new information about the ball's trajectory. I found that these predictions could be explained if people were generating new evidence over time, and integrating that with evidence from prior simulations to provide an updated belief about the future based on both past and present simulations. This suggests a process that our internal physics engine uses to provide us with regularly updated predictions about the future state of the world.

1.4 Conclusion

It is often argued the function of the brain is that of a 'prediction machine' (Clark, 2013), and that these predictions allow us to determine the consequences of our actions and choose our behavior accordingly (Grush, 2004). But in order to appropriately plan our actions, we need (relatively) accurate models that allow us to predict what will occur in the future. Despite the seeming ease with which we consider future events, forming accurate predictions requires both complex, structured knowledge about how the world functions, and the ability to integrate the uncertainty that we hold about the current state of the world into these models. This thesis demonstrates that the internal models we use for physical prediction accurately capture the rules of physics and robustly account for many sources of uncertainty inherent in the world, yet are flexible enough to provide a continuously updated view of what the future might hold.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

Bergen, B., & Wheeler, K. (2010). Grammatical aspect and mental simulation. *Brain and Language*, *112*(3), 150–158.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Brown, D. E. (1994). Facilitating conceptual change using analogies and explanatory models. *International Journal of Science Education*, *16*(2), 201–214.

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

Catrambone, R., Jones, C. M., Jonides, J., & Seifert, C. (1995). Reasoning about curvilinear motion: Using principles of analogy. *Memory and Cognition*, *23*(3), 368–373.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Cohn, A. G., & Hazarika, S. M. (2001). Qualitative spatial representation and reasoning: An overview. *Fundameta Informaticae*, *46*, 1–29.

Davis, E., Marcus, G. F., & Chen, A. (2013). Reasoning from radically incomplete information: The case of containers. *Advances in Cognitive Systems*, *2*, 273–288.

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

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2&3), 105–225.

Forbus, K. D. (1980). Spatial and qualitative aspects of reasoning about motion. In *AAAI*.

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

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

Forbus, K. D. (1994). *Qualitative spatial reasoning: Framework and frontiers*. Defense Technical Information Center.

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Towards a theoretical framework. In R. Michalski, J. Carbonaell, & T. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, p. 311).

Forbus, K. D., Nielsen, P., & Faltings, B. (1987). Qualitative kinematics: A framework. In *Proceedings of the International Joint Conference on AI* (pp. 430–436).

Frick, A., Huber, S., Reips, U.-D., & Krist, H. (2005). Task-specific knowledge of the law of pendulum motion in children and adults. *Swiss Journal of Psychology*, *64*(2), 103–114.

Friedman, S. F., & Forbus, K. D. (2009). Learning naive physics models and misconceptions. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

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

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy

Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(03), 377–396.

Halloun, I., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, *53*(11), 1056–1065.

Hamrick, J., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? Optimal mental simulation tracks problem difficulty. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 730–746.

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, *6*(2), 90–95.

Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, *22*(4), 411–430.

Howard, I. P. (1978). Recognition and knowledge of the water-level principle. *Perception*, *7*, 151–160.

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.

Kahneman, D., & Tversky, A. (1981). *The simulation heuristic.* (Tech. Rep.). DTIC Document.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory and Cognition*, *14*(4), 308–312.

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

Kerzel, D. (2002). The locus of “memory displacement” is at least partially perceptual: Effects of velocity, expectation, friction, memory averaging, and weight. *Perception & Psychophysics*, *64*(4), 680–692.

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

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

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

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

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

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.

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.

Michotte, A. (1963). *The perception of causality*. Basic Books.

Millington, I. (2010). *Game physics engine development: How to build a robust commercial-grade physics engine for your game*. Boca Raton, FL: Taylor and Francis.

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

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

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

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

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

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

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161.

White, P. A. (2012). The impetus theory in judgments about object motion: A new perspective. *Psychonomic Bulletin and Review*, *19*, 1007–1028.

Zwaan, R. A. (2003). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of learning and motivation*, *44*, 35–62.

2 Radically different physical intuitions in action and conception

Abstract

Does human behavior exploit deep and accurate knowledge about how the world works, or does it rely on shallow and often inaccurate heuristics? This fundamental question is rooted in a classic dichotomy in psychology: human intuitions about even simple scenarios are poor, yet their behaviors can exceed the capabilities of even the most advanced machines. Here we argue that this dichotomy is false: perceptually guided interactions with moving objects are often accurately calibrated to physical laws, while conceptual judgments about the same scenarios are inaccurate and variable. We asked participants to both interact with and draw the path of objects in ballistic motion and found that their interactions could be explained by accurate Newtonian inferences under uncertainty, while their drawings were idiosyncratic and often incorrect. Our results suggest that the contrast between rich and calibrated versus poor and idiosyncratic patterns of reasoning do not exist between domains of knowledge, but rather between domains of behavior.

2.1 Introduction

Humans function remarkably well in varied, uncertain environments, but psychological research has documented many dramatic failures of human reasoning: we can walk over precarious terrain and stack dishes in elaborate arrangements in a drying rack, but we have trouble explaining how gravity works in basic situations (Hecht & Bertamini, 2000; McCloskey, Washburn, & Felch, 1983). Such discrepancies between robust, effective behavior and dramatic errors in simple problems have fueled key debates in behavioral economics (Camerer, 1987), communication (Piantadosi, Tily, & Gibson, 2011), reasoning (Tversky & Kahneman, 1983), and most recently in the domain of intuitive physics (Marcus & Davis, 2013). Here we argue for a general resolution to such tensions: implicit, perceptually guided interactions with the world draw on knowledge that is adapted to our environment, while explicit conceptual tasks often rely on idiosyncratic and error-prone patterns of reasoning. More broadly, we suggest that these discrepancies are better explained by delineating between domains of behavior, rather than domains of knowledge.

People are often grossly inaccurate in simple intuitive physics judgments such as drawing future trajectories of an object in ballistic motion, dropped from a moving platform, or released from a circular ramp (Caramazza, McCloskey, & Green, 1981; McCloskey, Caramazza, & Green, 1980; McCloskey & Kohl, 1983; Proffitt & Gilden, 1989; Ranney, 1994), but when people predict trajectories of billiard balls, estimate properties of colliding objects, or judge the stability of towers, their physical reasoning is often very accurate and consistent with the principles of Newtonian mechanics (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2013). Prior literature has attempted to explain this discrepancy by suggesting that some human knowledge of physical principles is accurate, while other knowledge is erroneous (e.g., people can estimate the stability of stacked objects, but have erroneous conceptions of ballistic motion; Marcus & Davis, 2013). However, the studies that suggest human physics knowledge is inaccurate differ from those that suggest it is correct not only in the type of knowledge probed, but also in how they elicit behavior: studies that demonstrate failures of physical knowledge tend to rely on explicit queries about physical mechanisms (e.g., diSessa, 1993; McCloskey et al., 1980; Shanon, 1976), while studies that show accurate knowledge tend to require people to use physical principles as they would in their normal interactions with the environment (e.g., Battaglia et al., 2013). Differences might therefore be due to the types of tasks or behaviors rather than they types of knowledge required.

We test whether participants' behavior might differ between tasks that rely on the same physical principle using the classically-studied test-case of judging the ballistic trajectory of a ball released from a pendulum after the cord has been cut. In Experiment 1, participants each performed three distinct tasks: one explicit, conceptual task, drawing, and two interactive tasks, catching and releasing. The drawing task replicated a classic failure of intuitive physics in which participants were shown static pictures of pendulums and asked to draw the path that a ball would take if the cord were cut at various points (Caramazza et al., 1981). In the catching and releasing tasks participants observed a pendulum in motion and were asked either to position a bucket to catch the ball once the pendulum cord were cut by a 'knife' or release the ball from the pendulum so that it would be projected into a fixed bucket. All three of these tasks entailed solving the same physical problem – extrapolating the ballistic trajectory of a pendulum bob after the cord has been cut – so the systematic differences between human judgments in each task could arise only from the structure of the task itself, rather than differences in the underlying physical principles. We find that while peoples' explicit reasoning about such scenarios reveals erroneous and inconsistent behavior, their interactions with these systems are consistent with the 'noisy Newton' hypothesis (Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013) which holds that people form physical inferences by computing the effects of accurate Newtonian principles under uncertainty.

In Experiment 2, we evaluated whether the differences between the interactive and conceptual tasks was due to the tasks themselves, or differences in their respective stimulus presentations. In the interactive tasks, participants observed the pendulum in motion, but in the drawing task participants only observed a diagram of a pendulum on a sheet of paper. Because observing motion has been found to improve physical reasoning in conceptual tasks (Kaiser, Proffitt, & Anderson, 1985; Kaiser, Proffitt, Whelan, & Hecht, 1992), we investigated whether participants would demonstrate accurate physical knowledge on the drawing task only in the presence of pendulum motion. Although participants' judgments were different after observing motion, we find that these differences were driven by additional information about the velocity of the pendulum ball, but found no evidence that people used different, more accurate physical principles when they had access to richer stimulus information.

Together, these results suggest that peoples' knowledge of physics differs not by domain of knowledge, but rather depending on the domain of their behavior.

2.2 Experiment 1: Difference between interaction and conception

2.2.1 Methods

Participants

Thirty-five UC San Diego undergraduates (with normal or corrected vision) participated in this experiment for course credit. All participants gave informed consent to participate in accordance with guidelines set by the UC San Diego Institutional Review Board. Participants were collected over a span of two weeks, and the number of participants was deemed appropriate before analysis based on pilot work (Smith, Battaglia, & Vul, 2013). Three participants were removed from analysis because their performance indicated that they were often responding randomly (see Appendix, Figure A.1 for details).

Procedure

Participants performed three blocked tasks that involved predicting the ballistic trajectory of a ball released from a pendulum: interactive cutting, interactive releasing, and explicit drawing. Participants always performed the drawing task after the interactive tasks, but the order of the interactive tasks was randomized across participants.

In the interactive tasks, participants viewed a computer monitor from a distance of approximately 60 cm, which initially depicted a ball swinging from a cord, consistent with pendulum motion. At some point in time the cord would be cut and the ball would be released, thus entering ballistic motion. Beneath the pendulum there was always a bucket, and in every trial the participant's goal was to get the ball to drop into the bucket after being released. How they were allowed to interact with the scene differed between two tasks: catching and releasing. With the exception of one initial practice trial per task that familiarized participants with the task, the path of the falling ball was occluded in order to prevent participants from learning a simple relationship between the ball's release position and its landing position. At the end of each trial, participants were given binary feedback that indicated whether or not the ball successfully landed in the bucket (we found no evidence of learning from this feedback; see Appendix, Section A.2.3). A success would earn participants a point, and each participant's score was totaled across all trials; however, this score was used solely as motivation to engage with the task and did not influence any of our analyses.

Catching task. Participants were instructed to adjust the bucket's horizontal position using the mouse so that the ball would land in the bucket after being released. The release time was pre-determined and varied across trials. Participants were notified of where the cord would be cut by an icon of a knife, which would darken when the cord was about to be cut. The center of the bucket was recorded as the participant's judgment about where the ball would land (Figure 2.1, top).

Releasing task. The bucket was held fixed at a pre-determined position and participants were instructed to cut the pendulum cord by clicking the mouse at a time that would cause the ball to drop into the bucket. Cutting the cord was not allowed for an initial period of time randomly determined between 1.2-3.6s to avoid biases from participants who would attempt to cut the cord as quickly as possible. The time at which the cord was cut was recorded for each trial (Figure 2.1, bottom.)

For each of the two interactive tasks, participants repeated 48 different trials five times each in a randomized order. Trials were matched across tasks such that where the ball landed in a catching trial was the bucket position in the matched releasing trial. In the catching task, there were 16 distinct release times, crossed with three vertical distances between the nadir of the pendulum and position of the bucket either 20, 35 or 50% of

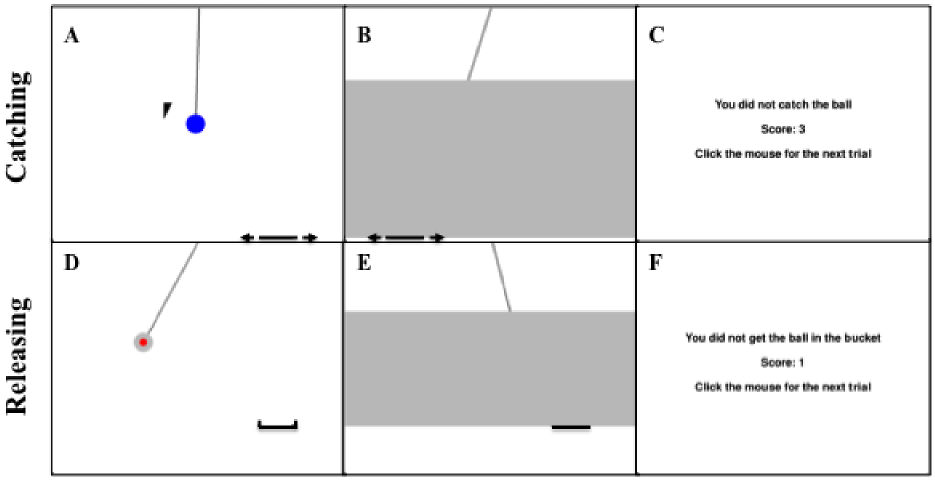

Catching
Releasing
A
B
C
You did not catch the ball
Score: 3
Click the mouse for the next trial
D
E
F
You did not get the ball in the bucket
Score: 1
Click the mouse for the next trial

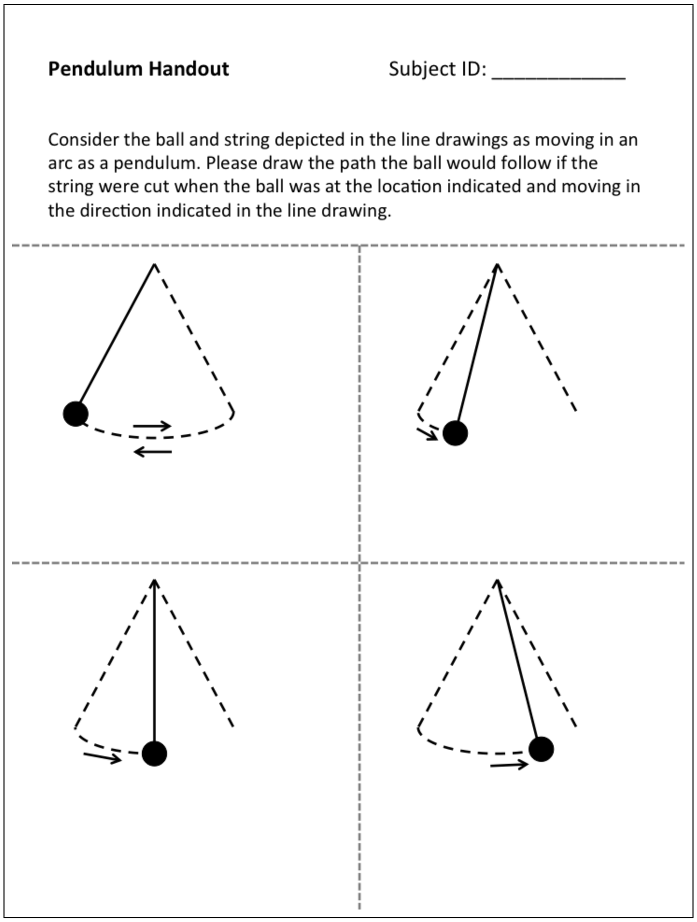
Figure 2.1: Diagram of trials in the catching (*top*) and releasing (*bottom*) trials. *Catching*: A. Participants observe a ball swinging on a pendulum and the 'knife' that will cut the cord. They move the bucket horizontally with the mouse. B. When the cord is cut, the trajectory of the ball is obscured. C. Binary feedback is provided after each trial. *Releasing*: D. Participants observe the ball on the pendulum. The coloring of the ball indicates a timer, such that once the red color is gone, participants can click the mouse to release the ball from the cord. E. The trajectory of the released ball is obscured. F. Binary feedback is provided.

the total screen height.

Both tasks and all trials used the same pendulum. This pendulum had a length of half of the screen, and reached a maximum angle of 35 degrees from vertical at its zenith. The period (2.5s) and force of gravity were set to obey Newtonian mechanics as if the pendulum were positioned at a depth of 6m from the participants. This depth was selected to conform to participants' general expectations about the natural period of the pendulum as seen on the 2D computer screen used in the experiment. To determine the motion of the pendulum, the cord was assumed to have negligible mass as compared to the ball, and so we could use simplified physical models to calculate the position of the pendulum at any point in time.

Drawing task. After the two interactive computer-based tasks, participants were given a two-page packet to fill out. On the first page was the *drawing* task – a set of four diagrams depicting pendulums at different points in their swings. Participants were asked to draw the path that the ball would take if the cord were cut at the time depicted in the diagram (Figure 2.2). On the second page was a brief survey that asked about the participant's number of prior physics courses and strategies used in the experiment. These questions were reviewed to check whether participants were responding based on surface-level features or using other strategies that did not involve prediction – however, we did not find evidence of this.

Drawings were classified into one of eleven patterns that were either drawing classifications from Caramazza et al. (1981) or were observed in a pilot experiment. Three undergraduate research assistants from both UCSD and MIT who were nave to the hypothesis performed this classification independently, and were told to match each participant's drawings to one of the given patterns as best as they were able, or rate the participant as *unclassifiable* if there were no matching pattern. A participant's drawing was considered matching a pattern if at least two of the three raters agreed; if all raters



Pendulum Handout Subject ID: ____________

Consider the ball and string depicted in the line drawings as moving in an arc as a pendulum. Please draw the path the ball would follow if the string were cut when the ball was at the location indicated and moving in the direction indicated in the line drawing.

Figure 2.2: Handout provided to participants for the drawing task. Instructions and stimuli were based on Caramazza et al. (1981).

disagreed, the participant's drawings were considered *unclassifiable*. There was high inter-rater reliability (Fleiss' $\kappa = 0.736$), and all three raters agreed on the classification for 23 of the 32 participants.

In addition to rating the drawings, we translated the drawing predictions to be comparable with the catching and releasing results by determining where participants would position the bucket in the catching task if their predictions were based on the drawing predictions. To perform this translation, we first fit either linear or quadratic functions through the ball and the lines participants drew.[1] The first author and three

[1] We could not simply determine where the drawn predictions crossed the line of each bucket height,

research assistants at UCSD marked on each drawing at least five points that accurately described the line (producing on average 54 points per drawing). We then fit two lines – linear and quadratic – through those points using least squares estimation. The quadratic fit was used for extrapolation unless either it had a positive quadratic term (implying the ball would move upwards), or the average distance between each point and the linear line was less than 1/8th cm more than the average distance from the quadratic line. In this way we allowed for curved drawings when appropriate but prevented inappropriate curvature that could bias results when the drawing itself was mostly linear. We then recorded where that extrapolated line crossed each of the three bucket heights, producing three 'pseudo-catching' results per drawing – this yielded 12 results per participant.

Models of physical reasoning

Even if people are using accurate physical principles to make predictions, uncertainty about the location or motion of objects can cause biases in the predictions themselves (Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013). Therefore, to test whether participants were using accurate physical principles, we designed a noisy Newton model to determine how people would behave if they were basing their predictions on Newtonian mechanics perturbed by uncertainty about the location of the pendulum and accumulated noise in the trajectory of the ball throughout extrapolation (Smith & Vul, 2013).

We split this model into two parts: the predictive *forward model*, and the *task action*. The predictive *forward model* describes how a trajectory is extrapolated, and thus the physical understanding presumed under the model – where the ball will go when cut from the cord; the forward model is a set of rules shared across the catching and releasing tasks. The *task action* determines how those predictions are used to position the

since many drawings did not extend that far or ended at the left or right side of the drawing area. Therefore, we used a common extrapolation technique for all drawings.

bucket or choose when to cut the cord, after incorporating noise/uncertainty from either accumulated prediction errors or motor control; these task actions differed across the two interactive tasks.

In addition to testing whether participants' predictions could be explained by noisy Newton reasoning, we also considered whether predictions could be better explained by alternate, non-physical reasoning. To test these non-physical accounts, we compared variants with different non-physical forward models against one another; the task actions, however, stayed constant between models.

Noisy Newton forward model. The noisy Newton physical forward model assumes that people have an accurate knowledge of the laws of ballistic dynamics. To predict where the ball will land, the model uses Newtonian ballistic motion equations to extrapolate the path of the ball given its position and velocity at the moment of release.

Although we assumed people have good knowledge of the laws underlying the pendulum system, we also assumed participants were uncertain about the distance of the pendulum in depth from the observer – a necessary assumption since the 2D image of a pendulum on a computer screen is not interpreted as a physical pendulum literally at the depth of the computer screen. There is a lawful relationship between pendulum period, cord length, and the force of gravity that people are sensitive to (Pittenger, 1985), and participants directly observe the period of the pendulum and are assumed to have a sense of realistic Earth gravity (McIntyre, Zago, Berthoz, & Lacquaniti, 2001). But because the pendulum was presented on a computer screen with no depth cues, people must infer how far behind the screen they expect the pendulum to be positioned, and therefore the length of the pendulum cord.

Non-physical forward models. Despite the body of literature studying the physical misconceptions that people hold, there is a dearth of formalized models about how people might understand ballistic motion; most research instead focused on conceptual descriptions of how gravity influences falling objects (Shanon, 1976) or how objects accelerate during their trajectory (Hecht & Bertamini, 2000). While Zago et al. (2004) suggest that people fail to account for gravitational acceleration in prediction, this only implies that non-physical models should predict that the ball travel in a straight line, but not the direction in which it is released. We therefore assumed that the extrapolations might be similar to those made on an explicit task, and formalized alternate forward models that could capture the same patterns participants made on the drawing task to test whether these patterns would be extended into the catching and releasing tasks. Each of the non-physical drawings could be captured by extrapolating the ball's path in a straight line at an angle away from the vertical; however, the calculation of that angle varied by model (see Figure 2.3).[2]

We tested three non-physical models: *angled, outward,* and *straight down*. The *angled* forward model calculated the ball angle as a piecewise linear function of the angle that the pendulum formed with its vertical at release, allowing for differences in trajectory on the downswing compared to the upswing. The *outward* model assumed that the ball would continue along the path of the cord, but allowed for the angle to shift upon release. Finally, the *straight down* model simply assumed that the ball would drop upon release.

Task actions. The forward models provide a single deterministic prediction of where the ball will travel given that the cord is cut in a certain position, but people must use this information to interact with the task, specifically choosing where to place the bucket in the catching task or when to cut the cord in the releasing task.

In the catching task, participants observed where the ball was cut from the cord, and were required to predict where it would land. This model captured human

[2] Non-linear extrapolated trajectories, such as adding a quadratic term to the path, would make these non-physical models equivalent to a physical model of a parabolic ballistic trajectory; thus only linear extrapolation paths are guaranteed to differ from physical extrapolation.

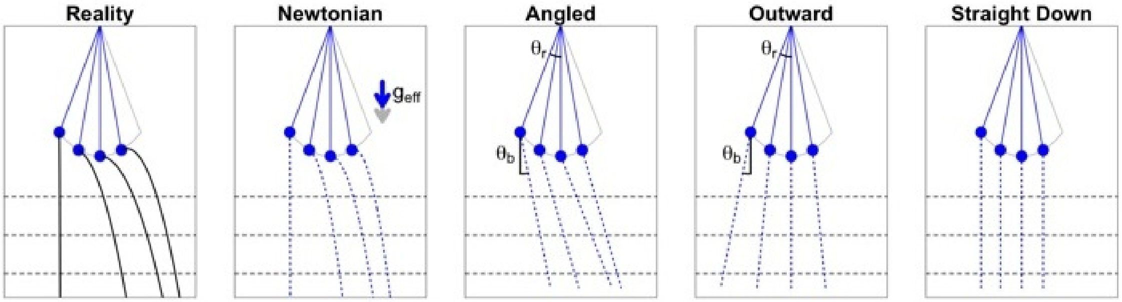


Figure 2.3: Diagrams of the forward model predictions for the path of the ball at four different cut points along the pendulum, using best fitting parameters. *Ground truth* is the path of the ball that was used to determine a successful catch in the experiment. See the Appendix, Section A.1 for mathematical formulations of each model.

performance by using the forward model to determine where the ball should go given its release, and assumed that this would be the average location that participants would place the bucket. However, participants' responses were variable, and the model must capture this. Noise in tasks that require catching a hidden falling object includes both predictive and motor uncertainty (Faisal & Wolpert, 2009), both of which were modeled together as Gaussian noise around the predicted position. Since prediction error accumulates throughout the path, the model's uncertainty increases linearly with the vertical distance between the bucket and the release height of the ball.

In the releasing task, participants needed to solve the inverse problem from the catching task: given a specific landing position, where in the pendulum period should the ball be when the cord is cut? We assume that people always have a reasonable sense of where the ball will go if released at each point in time. Assuming that motor errors are symmetric in time (Dawson, 1988), the optimal time to release the ball would be the middle of any contiguous period in which the ball would land in the bucket. If there were two contiguous periods of success,[3] we assumed that participants would be

[3] For instance, if the bucket were directly below the center of the pendulum, there are two periods when the ball could be released: when it is to the left of the bucket and traveling rightward, or when it is to the right of the bucket and traveling leftward.

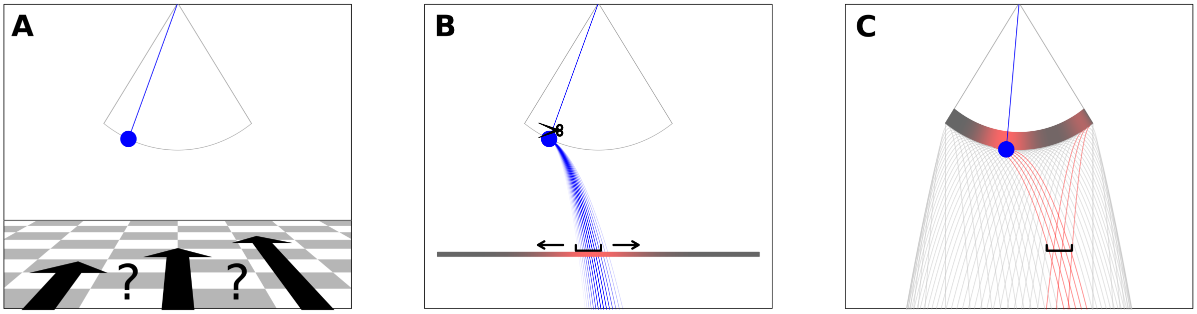


Figure 2.4: Illustration of the noisy Newton account of human judgments in the catching and releasing tasks. (A) Participants estimate the physical depth (and thus length) of the pendulum given its 2D projection and use this to guide both catching and releasing behavior. (B) In the catching task, participants see where the cord will be cut, generate noisy projections about where the trajectory of the ball will cross the plane of the bucket, and move the bucket into that region (red color mapping indicates higher probability of placing the bucket around that point). (C) In the releasing task, participants must choose when to cut the cord, so they project the ball's trajectory if released from different points spanning the pendulums' period, and choose a time to cut the cord that will make it probable that the ball lands in the fixed bucket given their motor timing error (red areas represent pendulum locations that are more likely to be selected as the release point).

probabilistically more likely to choose the release point with the shorter vertical distance between the ball and the bucket, as we share the assumption from the catching task that uncertainty accumulates over vertical distance. Finally, once the model chooses the time point that it aims to cut, its actual release time for a trial was perturbed by Gaussian noise to reflect the motor errors that people make.

2.2.2 Results

Consistency and accuracy of predictions

Participants' predictions in both of the interactive tasks were remarkably consistent with the noisy Newton model (catching: $r = 0.988$, releasing: $r = 0.998$, see Figure 2.5: Noisy Newton). Participants' drawings, however not only failed to capture veridical physical principles (only 6% of participants drew trajectories consistent with Newtonian physics; Figure 2.5: Drawing), but were also inconsistent from person to person (no more than 22% of participants were classifiable into a single category of response patterns). This drawing variability mirrors behavioral variability in other purely conceptual physical tasks (Caramazza et al., 1981; Kaiser, Proffitt, & McCloskey, 1985; Proffitt, Kaiser, & Whelan, 1990). If we extrapolate how participants would catch the ball based on their drawings, we find large variability in how well each participant's errors correlates with all other participants' errors (mean $r = 0.29$, 10-90% quantiles $=[-0.66, 0.89]$; see Figure S2). In contrast, individual participants' errors were much more consistently correlated with each other on both the catching (mean $r = 0.76$, 10-90% quantiles $=[0.47, 0.91]$) and releasing tasks (mean $r = 0.53$, 10-90% quantiles $=[0.33, 0.64]$, see Figure S2). These results suggest that behavior in interactive tasks is consistent across participants because it is calibrated to the physical world, while drawing behavior is not driven by physical principles and thus varies in ways that primarily reflect participants' idiosyncratic, explicit reasoning strategies.

The fit between human judgments and the noisy Newton model depends crucially on peoples' ability to reason effectively under uncertainty. Although participants' judgments were correlated with "ground truth" answers – responses under which the ball always landed in the center of the bucket (catching: $r = 0.969$, cutting: $r = 0.989$), judgments were systematically biased relative to ground truth (Figure 2.5: Ground Truth). Moreover, these systematic biases were different between the catching and releasing tasks (error correlation across matched trials: $r = 0.35$). These unique task biases are expected under the noisy Newton model because each task reflects different sources of uncertainty subjected to the same non-linear transformation via Newtonian kinematics; and indeed these systematic deviations of participants' judgments from the ground truth model matched the deviations of the noisy Newton model (catching: $r = 0.92$, releasing: $r = 0.93$). By capturing these systematic biases, the noisy Newton model correlated better

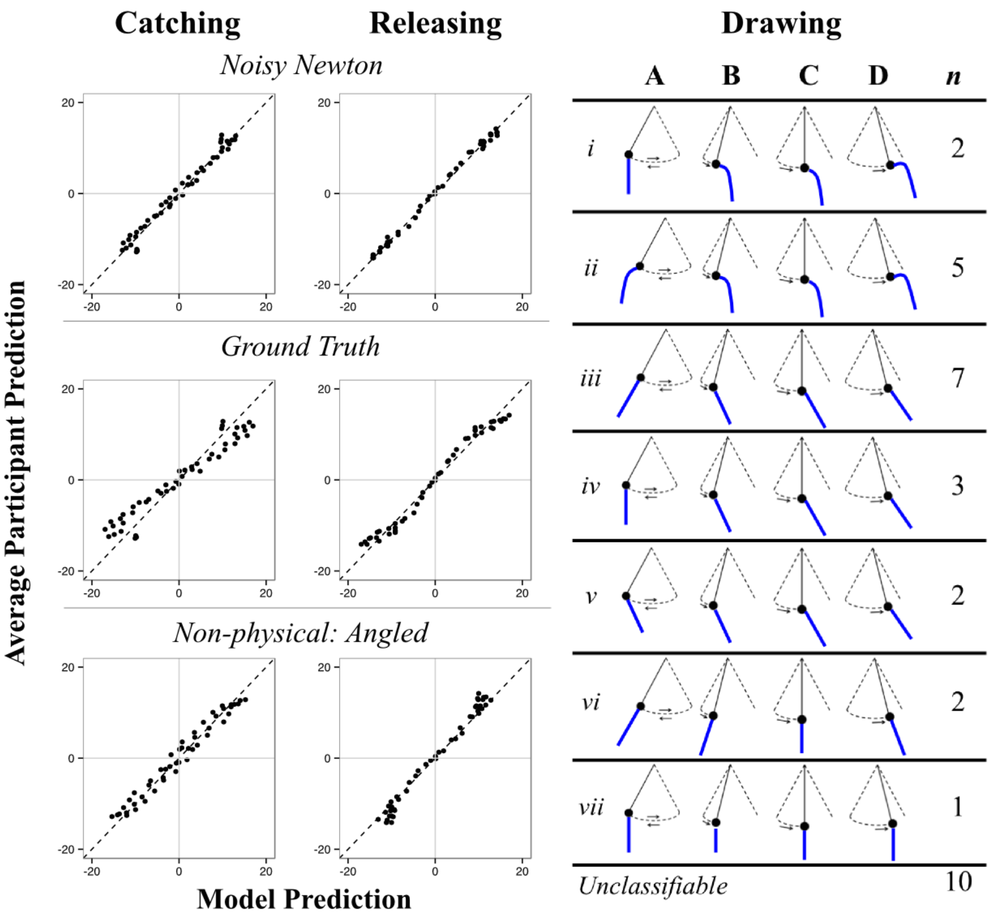


Figure 2.5: *Left*: The bias and variance of participants' average performance on the interactive (catching and releasing) tasks is better captured by the noisy Newton model than ground truth or the angled non-physical model. Each point represents one of 48 unique trials in either the catching or releasing tasks. On the x-axis are model predictions (in cm from the center of the screen) about the position of the bucket (catching) or the landing position of the ball if released at the predicted time (releasing), while the y-axis represents the average bucket (catching) or landing position (releasing) across all participants for that trial. *Right*: Classification of responses on the drawing task. Few (6%) participants drew accurate paths for all four diagrams (classification *i*), suggesting that most participants lack explicit knowledge of the physical principles underlying this task.

with participants' behavior than ground truth (catching: $z = 2.26$, $p = 0.02$, releasing: $z = 4.28$, $p < .001$). This suggests that apparent biases in implicit physical judgments reflect sophisticated patterns of probabilistic reasoning through an accurate physical model.

Behavior on the catching and releasing tasks also shows that people use accurate physical principles rather than incorrect, heuristic approximations. The noisy Newton model explained participants' catching and releasing responses better than any of the alternative forward models (angled: $\Delta BIC = 2,981$; outward: $\Delta BIC = 14,435$; straight-down: $\Delta BIC = 14,698$; see Figure 2.6), suggesting that participants were not typically using a non-Newtonian heuristic to extrapolate the ball's motion.

Individual physical knowledge

To test whether each participant was individually using Newtonian prediction, rather than such behavior arising only in the across-subject aggregate, we determined which of the noisy Newton and three heuristic models best described the behavior of each participant. Of the 32 participants, 28 (88%) were best fit by the Noisy Newton model, and 4 (12%) by non-physical models. Moreover, none of the four participants with non-Newtonian catching and releasing behavior had drawn extrapolated trajectories consistent with the heuristic model that best captured their interactive task behavior (see Table 2.1). These results suggest that the population does not contain subsets who have universally incorrect knowledge of physics across cognitive domains. Instead, when interacting with physical scenes, people share a common system of physical knowledge, calibrated with the world, while their deliberate judgments are guided by idiosyncratic and often non-Newtonian heuristics such as "conceptualiz[ing] an event's outcome in a representational context" (Kaiser, Proffitt, & McCloskey, 1985).

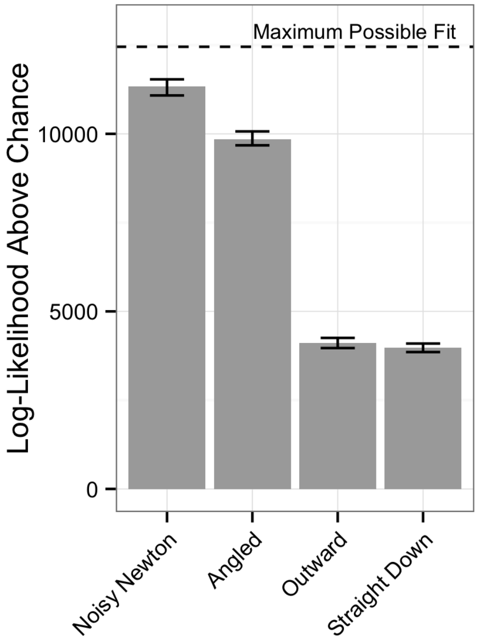


Figure 2.6: Fits of each model to human catching and releasing behavior. The noisy Newton model explains participants behavior better than any of the heuristic models. Log-likelihood above chance is the difference of the log-likelihoods of each of the models from the log-likelihood of a random response model. Maximum possible fit is the log-likelihood of predicting behavior as well as possible from the behavior of other participants. Error bars are 95% confidence intervals, calculated from 500 bootstrapped samples each.

Table 2.1: Individual best fitting model vs. classification on drawing task. Roman numerals refer to the drawing type classification from Figure 2.4. No participant was best fit by a non-physical model that could capture his or her drawing classification.

		Model Fit			
		Newtonian	**Angled**	**Outward**	**Down**
Drawing Task	**Newtonian** *(i, ii)*	6	1	0	0
	Angled *(iii, iv, v)*	12	0	0	0
	Outward *(vi)*	1	1	0	0
	Down *(vii)*	1	0	0	0
	Unclassified	8	1	1	0

2.3 Experiment 2: The impact of stimulus richness on physical knowledge

We found in Experiment 1 that interactive tasks tapped into relatively accurate models of physical reasoning, while participants relied on idiosyncratic, erroneous physical reasoning to solve conceptual tasks. However, the tasks in Experiment 1 differed not just in the way that we queried participants' knowledge, but also in the information available to participants to perform the tasks: in the catching and releasing tasks, participants observed the pendulum in motion, while in the drawing task participants were given a sheet of paper displaying a static pendulum. Prior work has suggested that viewing moving stimuli can produce more accurate physical judgments (Kaiser, Proffitt, & Anderson, 1985; Kaiser et al., 1992). However, these experiments introduce motion not just by showing the stimulus in motion before prediction (e.g., the pendulum in motion), but also query participants' judgments by showing various potential motion trajectories and ask participants to choose which is the most natural. Therefore it is not clear whether it is the initial motion information or the naturalness judgment that produces more accurate judgments. To tease these possibilities apart, we test how people perform the drawing task with a moving pendulum to determine whether participants with motion information would rely on more accurate physical principles. Although showing moving pendulums does change the predictions that people make on the drawing task, these changes are not due to people using more accurate physical principles but rather from making different inferences about the velocity of the ball at the moment the pendulum string is cut. Furthermore, participants who observed a moving pendulum produced more variable motion paths than those who judged only a static pendulum, suggesting that motion may make these explicit judgments less accurate.

2.3.1 Methods

Participants

Sixty-seven UC San Diego undergraduates (with normal or corrected vision) participated in this experiment as part of a set of experiments for course credit. All participants gave informed consent to participate in accordance with guidelines set by the UC San Diego Institutional Review Board. We collected data until we had approximately twice the number of participants from the original task. Participants were randomly assigned to the *Motion* or *Static* conditions, resulting in 33 participants in the Motion condition and 34 participants in the Static condition.

Procedure

Participants were instructed that they would need to judge the path of a ball that is cut from a pendulum, and that they would indicate the ball's predicted path by clicking and dragging the mouse. Participants in the *Motion* condition observed the pendulum make one full swing then swing to the point where the string would be cut, while participants in the *Static* condition observed only the final position of the pendulum as the string is cut; therefore participants in both conditions observed identical images immediately before being asked to respond. The pendulum used here was identical to the pendulum used in the *Catching* and *Releasing* tasks, with the same arc and, for the *Motion* condition, the same period.

Participants indicated their predictions by clicking and dragging along the path they believed the ball would travel. To ensure that we captured paths of appropriate length, these paths were required to (a) start from within the image of the ball and (b) terminate within 10% of the edge of the lower half of the screen; if the path did not meet these criteria, participants were notified and asked to draw the path again. Finally, participants would be asked to either confirm their path, or click a ‘Try Again’ button to re-draw it (see Figure 2.7: top).

All participants drew their predictions for the same four release points measured in the *Drawing* part of Experiment 1; the order of presentation was randomized across participants. For each drawing, we recorded each point along which participants dragged the mouse, measured every 20ms, from which we could reproduce the drawn path.

Rating

As with the *Drawing* task of Experiment 1, we asked three undergraduate raters from UCSD to classify each participant’s drawings. Because we hoped to test how judgments varied in detail, we asked the raters to judge the predictions individually by stimulus, rather than the pattern of predictions across all four stimuli. Raters classified each drawing into one of six types (see Figure 2.7: bottom), or judged an individual drawing to be *unclassifiable*. Raters were blind to which participant created each stimulus and to whether they were in the *Motion* or *Static* condition.

Inter-rater reliability was lower than the reliability from Experiment 1 (Fleiss’ $\kappa = 0.596$), but this effect was driven by one rater who had a higher threshold for classifying drawings (rating 42% of drawings as unclassifiable). Reliability where this rater classified drawings was very high (Fleiss’ $\kappa = 0.826$), and on the stimuli she determined to be unclassifiable the other two raters agreed on a classification 79% of the time. Similar to Experiment 1, we classified each drawing as the majority classification of the raters, but if all three raters disagreed, we noted the drawing as unclassifiable (this was only true of 5% of the drawings).

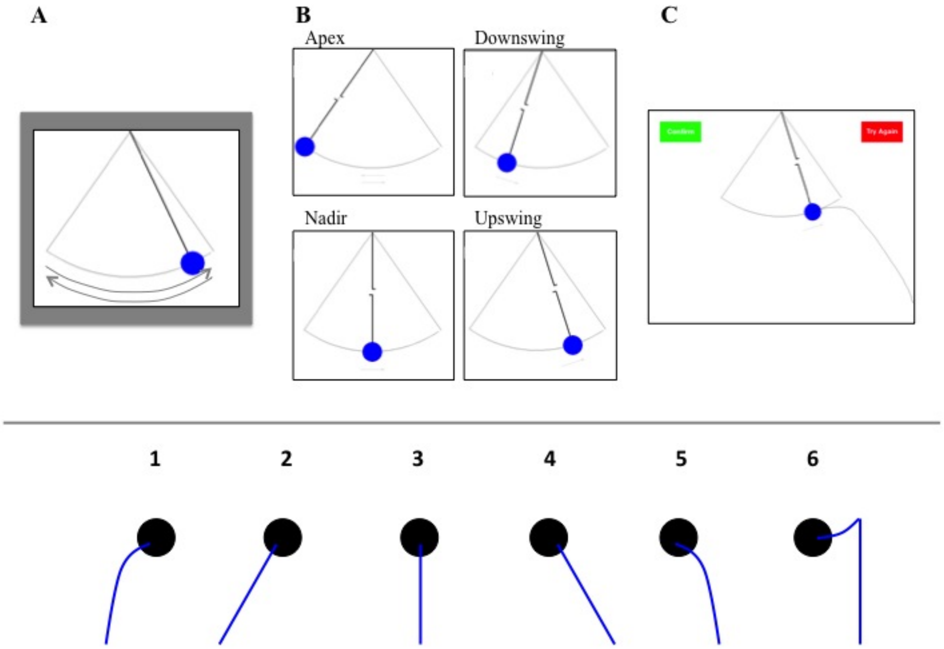


Figure 2.7: *Top*: Diagram of a trial. A: participants in the *Motion* condition only observed the pendulum swing through one full period, then swing to the final position. B: participants in both conditions would observe a static image of the pendulum string cut at one of four positions. C: participants click and drag the mouse to indicate their predictions for the balls motion. *Bottom*: The six potential paths raters could classify each drawing as (not including *unclassified*). All of the patterns from Experiment 1 or Caramazza et al. (1981) could be recreated from a combination of these path types.

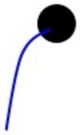
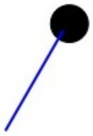
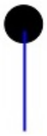
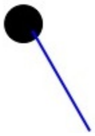
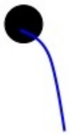
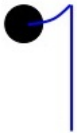
2.3.2 Results

We first tested whether there was evidence of differences in participants' predictions due to motion evidence for each pendulum cut point. If motion information does not affect physical reasoning, then we should expect no difference between participants' predictions in the *Motion* and *Static* conditions. On the other hand, if motion information causes people to use accurate models of physics, then participants in the *Motion* condition should make different predictions from those in the *Static* condition, and should draw more curved paths to indicate the appropriate influence of gravity on the ball's ballistic trajectory.

We did find evidence that participants' drawings differed between the two conditions in for both the Apex ($\chi^2 = 13.4, p_{sim} = 0.014$) and the Nadir ($\chi^2 = 10.7, p_{sim} = 0.035$) pendulums, but not in the Downswing ($\chi^2 = 2.4, p_{sim} = 0.71$) or Upswing ($\chi^2 = 8.3, p_{sim} = 0.14$) stimuli. The differences in Apex predictions appear to be driven by participants with motion information believing that the ball retains leftward velocity, while participants without motion information tend to believe the ball will drop (the correct answer) or travel to the right. The difference in Nadir predictions are driven by participants without motion information indicating that the ball will drop straight down, while participants with motion information realize that the ball retains horizontal velocity (see Table 2).

Although there is evidence that motion does influence peoples' predictions, there is no evidence that it causes them to use more accurate physical principles for those predictions. For the Downswing, Nadir, and Upswing stimuli, there was no evidence that participants in either condition drew the correct ball path at different rates (path 5 from Figure 7; all $\chi^2 < 0.5$, all $p_{sim} > 0.5$). There was a difference in accuracy with the Apex condition, but it was participants in the Static condition who were more likely to be correct (24% vs. 6%; $\chi^2 = 4.4, p_{sim} = 0.043$). Thus pendulum motion provides

Table 2.2: Classification of participants drawings, split by pendulum cut point and experimental condition. The veridical response was 5 in all cases except the Apex, where the veridical response was 3. Patterns of responses between the *Static* and *Motion* conditions differ in the Apex and Nadir scenarios based on differences in how participants interpret the ball's velocity, but there is no evidence that the physical principles used differ between conditions.

		1 	**2** 	**3** 	**4** 	**5** 	**6** 	**Unclassified**
Apex	Static	5	1	8	10	1	0	8
	Motion	10	9	2	5	1	0	7
Downswing	Static	0	2	2	18	7	0	6
	Motion	0	2	0	19	7	0	4
Nadir	Static	1	1	13	7	7	0	8
	Motion	0	0	5	16	5	0	4
Upswing	Static	0	7	1	8	8	3	6
	Motion	0	0	2	10	10	5	6

different information about the ball's velocity, but this information can be misleading (e.g., indicating the ball retains velocity at the apex) and does not cause people to produce more correct parabolic paths.

Consistency of Static and Motion predictions

We extrapolated the drawings in the same way as Experiment 1 as a separate test of how consistent the *Motion* and *Static* predictions were (see Section 2.2.1 – Drawing Task).[4] If participants were using qualitatively different reasoning between the two conditions, then individual predictions from the *Motion* condition should correlate better with the average of others from the *Motion* condition than with those from the *Static* condition, and vice versa.

[4] Because we captured points along the drawn line as part of the task, we did not have third parties mark each drawing, but the technique for extrapolating lines from the drawn points was identical.

Similar to Experiment 1, participants drawing errors were not well correlated with the average errors from all other participants and were extremely variable (mean $r = 0.29$, 10-90% quantiles $=[-0.62, 0.89]$). However, this did vary as a function of condition: extrapolated drawings from the Static condition were more correlated with other Static extrapolations than Motion extrapolations (Static: mean $r = 0.50$, 10-90% quantiles $=[-0.13, 0.89]$; Motion: mean $r = 0.02$, 10-90% quantiles $=[-0.39, 0.47]$), while the Motion extrapolations were somewhat more similar to other Motion extrapolations than Static (Static: mean $r = 0.07$, 10-90% quantiles $=[-0.78, 0.94]$; Motion: mean $r = 0.16$, 10-90% quantiles $=[-0.40, 0.65]$).

However, this effect was driven almost exclusively by differences in prediction for the Nadir cut point; excluding this stimulus, participants in the *Static* condition were equally well correlated with the average extrapolations in both conditions (Static: mean $r = 0.45$, 10-90% quantiles $=[-0.49, 0.91]$; Motion: mean $r = 0.42$, 10-90% quantiles $=[-0.12, 0.87]$), as were participants in the *Motion* condition (Static: mean $r = 0.17$, 10-90% quantiles $=[-0.80, 0.95]$; Motion: mean $r = 0.22$, 10-90% quantiles $=[-0.67, 0.81]$).

This provides further evidence that seeing the pendulum in motion provides additional information about the velocity of the ball at the moment that the string is cut, but does not make the physical principles that underlie the motion extrapolation any more accurate. In fact, the lower correlations and higher variability of the *Motion* extrapolations suggests that if anything, observing motion makes predictions even more variable and therefore less veridical.

2.4 Discussion

In two experiments we asked people to make physical judgments in several different tasks, all of which depended on identical underlying physical principles. Participants were surprisingly accurate for perceptually guided interactions, but idiosyncratic and inaccurate for explicit, conceptually based responses. In Experiment 1, participants demonstrated accurate use of physical principles about ballistic motion when interacting with a scene, but those same participants were often erroneous when asked to conceptualize the same principle by drawing an object's trajectory. In Experiment 2, participants continued to use erroneous physical principles on conceptual tasks, even with richer, less abstract stimulus information.

These findings mirror a broader pattern of results in the psychological literature: peoples' behavior differs between conceptual and perceptual-motor tasks (Chen, Ross, & Murphy, 2014; Glaser, Trommershäuser, Mamassian, & Maloney, 2012; Wu, Delgado, & Maloney, 2009). Some behavior, especially in lower level perceptual and motor domains, is near optimal given the information and processing constraints associated with a particular task (Griffiths & Tenenbaum, 2006; Stocker & Simoncelli, 2006; Trommershäuser, Landy, & Maloney, 2006; Wolpert, Ghahramani, & Jordan, 1995) while other behavior, especially in higher level cognition, is subject to gross biases and errors (McCloskey et al., 1980; Tversky & Kahneman, 1983). This dichotomy has driven debates as to whether cognition is generally rational (Anderson, 1990; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) or whether it is based on a set of ad hoc heuristics (Gigerenzer & Gaissmaier, 2011; Marcus & Davis, 2013). Our results suggest an alternate contrast: Everyday behavior is calibrated and sensible by combining uncertainty with rich, accurate knowledge about how the world works; on the other hand, deliberate reasoning cannot access such rich world knowledge and instead relies on explicit and potentially faulty information. A basketball player might weave past opponents to score a spectacular basket but not be able to explain what he is about to do, a child's ability to draw a cat is not related to her ability to recognize one, and most people can speak coherently without explicit knowledge of how to conjugate verbs.

Our results indicate that the contrast between effortless, calibrated actions and idiosyncratic, error-prone reasoning does not exist between domains of knowledge, but rather between domains of behavior. If people have multiple, and sometimes inconsistent, systems of knowledge, debates about human rationality can be refocused on explaining the structure and organization of our world knowledge, how it guides different types of behavior, and how it arises over experience and evolution. Rather than argue whether people do or do not understand certain physical principles, for instance, we should study how people develop the physical understanding needed to interact with the world, how people reason explicitly about physics, and how those two systems interact to determine behavior in different situations. By accounting for these different systems of knowledge, we can move beyond simply debating whether or not the human mind contains a specific concept and instead study how that concept might be incorporated in different systems of knowledge and deployed differently between domains of behavior.

Acknowledgments

The authors would like to thank Frank Jäkel, Nancy Kanwisher, Joshua Tenenbaum, Gary Marcus, and Ernest Davis for their helpful discussion and comments. KAS and EV were supported by NSF CPS grant 1239323, and a UCSD Interdisciplinary Collaboratories grant. KAS, EV and PB were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023.

Chapter 2, in full, is currently under review for publication of the material in

Cognition. Smith, Kevin A; Battaglia, Peter W; Vul, Edward. The thesis author was the primary investigator and author of this material.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

Camerer, C. F. (1987). Do biases in probability judgments matter in markets? Experimental evidence. *The American Economic Review*, *77*(5), 981–997.

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

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General*, *143*(1), 227–246.

Dawson, M. R. W. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments & Computers*, *20*(1), 54–57.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2&3), 105–225.

Faisal, A. A., & Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology*, *101*, 1901–1912.

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

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Glaser, C., Trommershäuser, J., Mamassian, P., & Maloney, L. T. (2012). Comparison of the distortion of probability information in decision under risk and an equivalent visual task. *Psychological Science*, *23*(4), 419–426.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal prediction in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 730–746.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 795–803.

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

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

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

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

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.

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.

McIntyre, J., Zago, M., Berthoz, A., & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience*, *4*(7), 693–694.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pittenger, J. B. (1985). Estimation of pendulum length from information in motion. *Perception*, *14*, 247–256.

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

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

Ranney, M. (1994). Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke and Breedin. *Memory & Cognition*, *22*(4), 494–502.

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

Shanon, B. (1976). Aristotelianism, Newtonianism and the physics of the layman. *Perception*, *5*(2), 241–243.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

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

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Trommershäuser, J., Landy, M. S., & Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science*, *17*(11), 981–988.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences*, *106*(15), 6088–6093.

Zago, M., Bosco, G., Maffei, V., Iosa, M., Ivanenko, Y. P., & Lacquaniti, F. (2004). Internal models of target motion: Expected dynamics overrides measured kinematics in timing manual interceptions. *Journal of Neurophysiology*, *91*, 1620–1634.

3 Sources of uncertainty in intuitive physics

Abstract

Recent work suggests that people predict how objects interact in a manner consistent with Newtonian physics, but with additional uncertainty. However, the sources of uncertainty have not been examined. Here we measure perceptual noise in initial conditions and stochasticity in the physical model used to make predictions. Participants predicted the trajectory of a moving object through occluded motion and bounces, and we compared their behavior to an ideal observer model. We found that human judgments cannot be captured by simple heuristics, and must incorporate noisy dynamics. Moreover, these judgments are biased consistently with a prior expectation on object destinations, suggesting that people use simple expectations about outcomes to compensate for uncertainty about their physical models.

3.1 Introduction

Predicting how the world will unfold is key to our survival and ability to function on a daily basis. When we throw a ball, cross a busy street, or catch a pen about to fall off of a desk, we must foresee the future physical state of the world to plan our actions. The cognitive mechanisms that help us make these predictions have been termed ‘intuitive physics’ models.

Although human performance in physical prediction tasks tends to approximate real-world (Newtonian) physics, it does not match exactly: people make systematic prediction errors. While this has been taken as evidence that human models of intuitive physics are non-Newtonian (e.g., McCloskey, 1983), more recently human behavior has been explained by intuitive Newtonian physics models under uncertainty. On this account, human predictions deviate from Newtonian mechanics because of stochastic error – uncertainty about the initial positions or velocity of objects propagates through the non-linear physical model and causes variability and bias in final judgments. For instance, human predictions about the stability of a tower of blocks or the most likely direction for that tower to fall are consistent with a purely Newtonian model of physics with a small amount of uncertainty in the initial positions of the constituent blocks (Hamrick, Battaglia, & Tenenbaum, 2011). Similar models of physics with perceptual noise have been used to explain relative mass judgments in collisions (Sanborn, Mansinghka, & Griffiths, 2009) and infants’ expectations for object movement (Teglas et al., 2011).

There are numerous ways in which uncertainty can be introduced into intuitive physical reasoning. We broadly classify these into two categories: perceptual uncertainty and uncertainty about dynamics. Perceptual uncertainty arises because initial measurements of the location and velocity of objects is imperfect; this initial noise propagates through the model. Uncertainty about dynamics reflects noise in the physical model itself. Real object movement and collisions are perfectly deterministic only in an idealized system; in the world, objects can deviate from their ideal path because of multiple, unknowable interactions with the environment (e.g., a ball rolling across gravel will not move in a straight line). Stochastic dynamics could thus reflect such environmental uncertainty.

Our goal is to disentangle the influence of initial noisy percepts and noisy physics on human predictions of object dynamics. We compared human behavior in a simple physical prediction task to a stochastic physics model with parameters reflecting the different types of uncertainty.

3.2 Stochastic physics model

We designed a model to replicate stochastic physics in a simple environment: a ball bouncing around a two-dimensional box. We based this model on idealized mechanics, but also incorporated the two sources of uncertainty: we added noise to the initial position and velocity to capture perceptual uncertainty, while dynamic uncertainty was captured by jitter in object movement over time, and variability in bounce angles.

3.2.1 Uncertainty parameters

The model was based on a simple two-dimensional physics engine customized to add our sources of uncertainty. As physical uncertainty goes to zero, this model reduces to laws from idealized mechanics: the ball would continue to move in a straight line at a constant velocity until it hit a wall, at which point it would bounce elastically and with angle of incidence equal to the angle of exit. Uncertainty was captured using four parameters, two for the perceptual error, and two for the stochastic error.

Perceptual uncertainty

At the start of the simulation, the ball's position and velocity were based on where the ball would be in a perfectly deterministic simulation, but with noise added. Position was perturbed by isotropic two-dimensional Gaussian noise parameterized by standard deviation, σ_p. Noise in velocity direction was captured in a von Mises (circular normal)

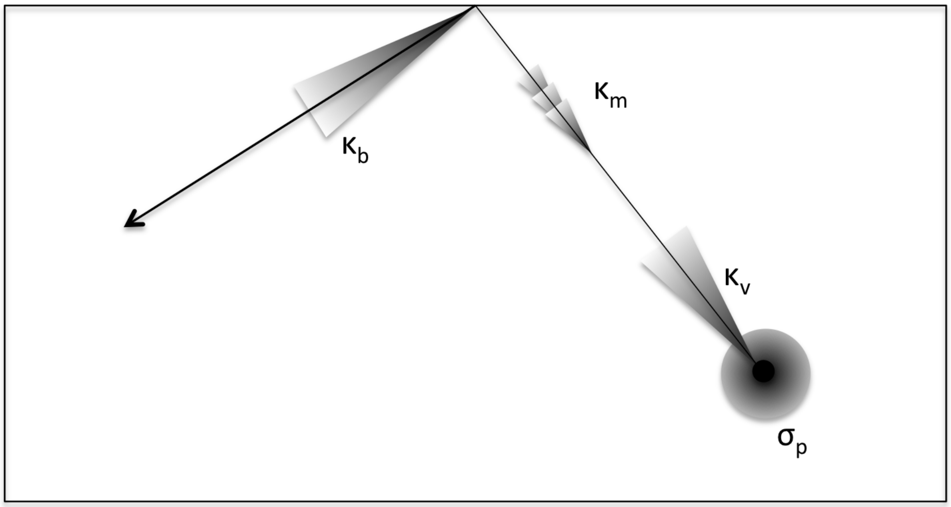


Figure 3.1: Sources of uncertainty in the stochastic physics model. Each parameter refers to a different source of noise: position noise (σ_p), velocity direction noise (κ_v), ongoing movement 'jitter' (κ_m), and noise added with a bounce (κ_b).

distribution on direction of motion, parameterized by concentration (inverse variance) κ_v. We did not consider uncertainty in the speed of the ball, as this would only affect the timing of the ball's movement but not its destination, which is the prediction we aim to capture.

Dynamic uncertainty

Noise was added during the simulation in two ways. First, at each time step (1000/sec), the direction of the ball was 'jittered' by adjusting its direction using a von Mises distribution with the concentration parameter κ_m. In addition, noise was added during each bounce by assuming that the angle the ball bounced off of the wall was defined by a von Mises distribution centered on the angle of incidence with a concentration parameter κ_b.

3.3 Experiment

We aimed to test model predictions against human data and to estimate uncertainty parameters in intuitive dynamics. In this experiment, subjects predicted the trajectory of a ball in a two-dimensional environment on a computer screen. This was done in a 'Pong' game where participants tried to catch the ball with a paddle. Crucially, we occluded the latter part of the ball's movement, so that successful prediction of the final position required the mental simulation of the object trajectory. We could estimate the final position predicted by our stochastic physics model with different parameters, and thus compare human behavior to model predictions under varying types and degrees of uncertainty.

In this experiment we parametrically varied both the distance the ball would travel[1] and the number of bounces off of walls while occluded. If intuitive dynamics models are deterministic, then the number of bounces will have no effect on human predictions. The distance manipulation was designed to tease apart the contributions of perceptual uncertainty about velocity and dynamic velocity noise.

3.3.1 Methods

52 UCSD undergraduates (with normal or corrected vision) participated in the experiment for course credit.

Subjects used a computer mouse to control the vertical position of an on-screen 'paddle' to catch a moving ball. The ball moved according to the deterministic physics underlying the stochastic physics model. Both the paddle and the ball were confined to a 1200 by 900 pixel area in the center of the screen. Each trial began with a display of only the paddle, which subjects could move up and down. The paddle was 100 pixels in

[1] Because the ball always moved at a constant velocity, distance was proportional to duration of occlusion.

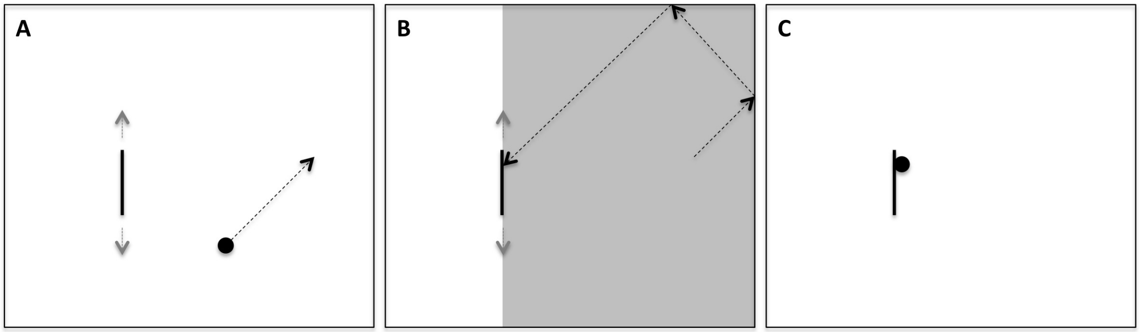


Figure 3.2: Diagram of a trial. (A) The ball moves unoccluded in a straight line. (B) Once the field is occluded, the ball continues to move and the subject must predict where it will end. (C) The trial ends once the ball is either caught or passes the plane of the paddle.

height and was centered on the vertical position of the mouse before each trial. A mouse click triggered the start of a trial. A ball would then appear on the screen, moving at a constant velocity of 600 pixels/second. After the ball moved 400 pixels (667ms), a grey rectangle would occlude the portion of the screen containing the ball (Figure 3.2). During this period, the ball would continue to move, bouncing perfectly elastically off of the edges of the field, but would not be visible. Once the subjects caught the ball with the paddle, or the ball broke the plane of the paddle, the trial would end and the occluder would be removed, showing whether (and by how far) the subject missed the ball. Upon clicking the mouse, the screen would clear and reset for the next trial. The number of balls caught by the subject was always displayed in the upper right corner as a motivation to perform well.

Subjects were given 648 trials throughout the experiment. These 648 trials were identical for all subjects, but presented in a randomized order. Each trial had a particular ball trajectory, generated by one of nine conditions. The nine trajectory conditions crossed the distance the ball travelled while occluded (600, 800, or 1000 pixels) with the number of bounces (0, 1, or 2); there were 72 trials of each condition. The specific path for each trial was generated prior to the experiment subject to the constraints of the condition and the constraint that the final position was not in the top 20% or bottom 20% of the enclosed area to avoid bias due to positioning the paddle at the ends of the screen.

Before starting the experiment, subjects were given seven trials without the occluder to demonstrate how the ball would move, then six practice trials with the occluder.

For each trial, we recorded the position of the midpoint of the paddle once the ball was caught or moved past the paddle. From this measure we could calculate, for each trial, (a) the average expected position of the ball, and (b) the variance of predictions around that expectation.

3.3.2 Subject performance

Accuracy

Subjects caught the ball on 43.8% of all trials. Individual subject accuracies varied between 25.6% and 63.7% (chance was 11%). Accuracy also varied by trial condition: subjects were most accurate in the shortest, no bounce condition (69%) and least accurate in the longest, two-bounce condition (32%).

Accuracy improved slightly over time, increasing from 42.7% in the first half of trials to 44.9% on the second half ($\chi^2(1) = 15.9, p < 0.001$). However, because this was a small effect, and because in a logistic model predicting accuracy, trial order did not interact with either distance ($\chi^2(2) = 0.72, p = 0.70$) or number of bounces ($\chi^2(2) = 4.18, p = 0.12$), we do not try to account for this change.

Expected positions

In addition to decreasing accuracy, subjects also showed increasing bias in average predictions as the distance or number of bounces increased. The mean final position of

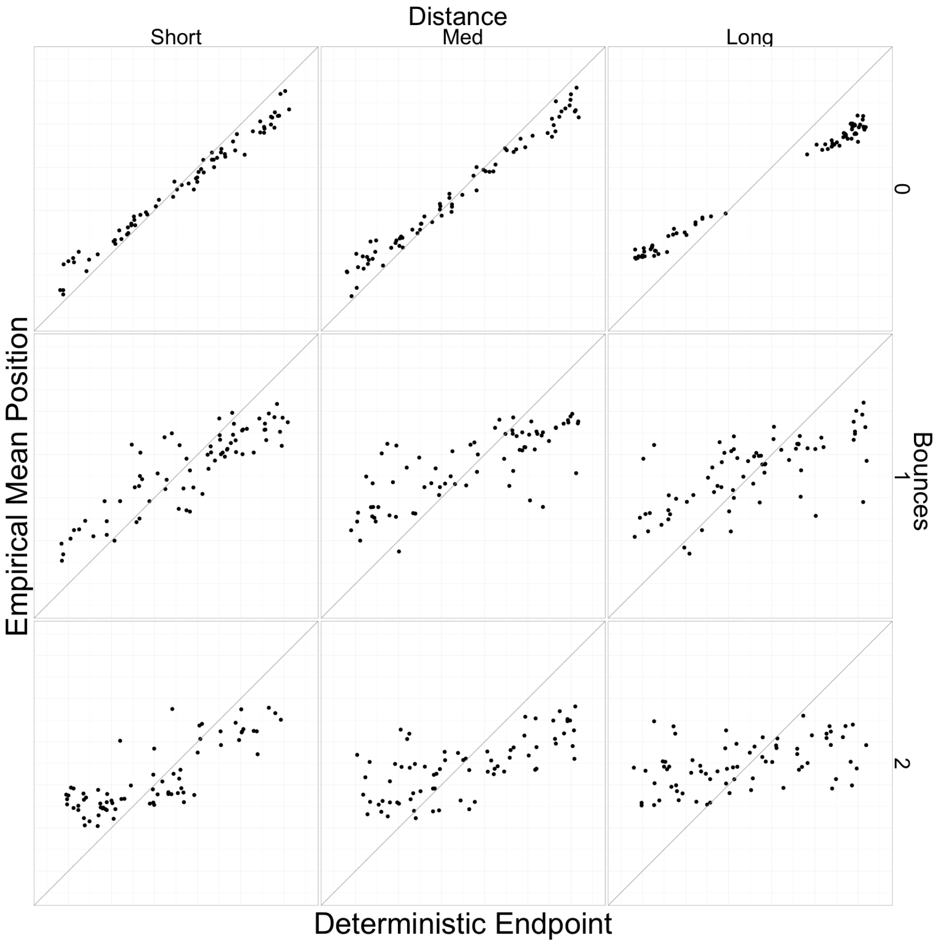


Figure 3.3: Mean predicted paddle position versus path endpoint using deterministic physics as a function of trial condition. Each point represents a separate trial.

Table 3.1: Percent of distance ‘shifted’ from actual end ball position towards center by trial condition

		Distance		
		600	800	1000
Bounces	0	24%	44%	53%
	1	23%	60%	70%
	2	41%	63%	84%

Table 3.2: Average standard deviation (in pixels) of responses within a trial by condition

		Distance		
		600	800	1000
Bounces	0	65	76	94
	1	111	115	114
	2	115	111	121

the paddle for each trial shifted towards the center as compared to the final ball position (see Figure 3.3). The magnitude of this bias toward the center of the screen increased as either distance or number of bounces increased.

Variance of responses

The variability of subjects’ responses around the mean also increased with distance and bounces, but only up to a ceiling – well below the maximum possible spread – once subjects had to take into account even one bounce (see Table 3.2).

3.4 Model application

The coarse results suggest that prediction error and variability increases with distance or number of bounces. But they do not indicate which sources of uncertainty contribute to intuitive physics predictions, nor do they explain why some trials within the same condition produce greater bias and variability than others.

We aimed to tease these factors apart via our model of stochastic physics. By finding the set of uncertainty parameters that best fits the empirical data, we can compare the relative contribution of the perceptual uncertainty parameters to the dynamic uncertainty parameters. A good model should capture trial-level differences in subjects' performance, and explain trial difficulty based on the interplay of different sources of uncertainty.

3.4.1 Simulation

We replicated the experimental task in the stochastic physics model, simulating the same 648 trials. To mirror this task, each simulation started at the point of occlusion (when subjects could no longer visually track the ball and must predict its path) and ended when the simulated ball crossed the plane of the paddle. On each simulation, we measured the position of the simulated ball along that plane. Because there is no analytic form of the probability distribution over possible trajectories, we simulated each trial 500 times, thus estimating the predictive distribution for each trial via sampling.

No reasonable set of uncertainty parameters produced mean estimates of the final position of the ball that were systematically shifted toward the center like the empirical data; as long as Newtonian physics underlies the model, averaging over all simulation paths, the mean ending position will be close to the actual endpoint for most trials, regardless of the uncertainty parameters chosen.[2] Since the magnitude of the center bias scaled with distance and number of bounces, we suspected that subjects were incorporating a prior on final position, producing a center bias proportional to the uncertainty in their physics-based predictions. People therefore appear to incorporate

[2] If the ball ended close to a bounding wall, the distribution of simulated end positions was skewed away from the wall (because of simulated bounces). However, the average end position tracked the actual endpoint with considerable fidelity ($r = 0.95$).

prior expectations with their intuitive physics models.

We treated this bias as a simple Gaussian prior on the final ball position centered on the middle of the screen, with standard deviation as a free parameter (σ_{prior}). One value of this parameter was used for all trials and conditions.

The final distribution of predictions for each trial was calculated by combining the center-prior with the distribution of predicted positions simulated by the stochastic physics engine. We treated the distribution of predicted positions as a Gaussian and calculated their mean and standard deviation. We could then calculate the mean and standard deviation of the posterior distribution using Bayesian cue combination (e.g., Ernst & Banks, 2002):

$$\sigma^2_{post} = \left(\frac{1}{\sigma^2_{prior}} + \frac{1}{\sigma^2_{sim}} \right)^{-1} \tag{3.1}$$

$$\mu_{post} = \left(\frac{x_{center}}{\sigma^2_{prior}} + \frac{\mu_{sim}}{\sigma^2_{sim}} \right) * \sigma^2_{post} \tag{3.2}$$

Using these equations, trials with greater simulation variance will be more affected by the prior, and will shift further towards the screen center. Thus, the model can account for the center-bias in a manner sensitive to prediction uncertainty.

We found the maximum likelihood parameters to fit three quarters of the data (with an equal number of trials from each of the distance by bounce conditions).[3] We also fit two other models: one with only perceptual uncertainty and prior parameters, and one with only dynamic uncertainty and prior parameters. We compared these models based on the likelihood of the quarter of the remaining (cross-validation) data.

[3] Numerical optimization techniques can find local minima, so we used multiple starting points and grid search across 1,600 sets of parameters to ensure we were finding the global minimum.

3.4.2 Model results

Model comparison

We designed the stochastic physics model to investigate how various sources of uncertainty contribute to intuitive physics. Thus we compared the model with both dynamic and perceptual uncertainty to the two nested models with either dynamic or perceptual uncertainty parameters alone to determine which sets of parameters were necessary to best explain the data.

In addition, we tested how well any of the stochastic models captures human behavior by comparing them to a 'heuristic oracle' model with different parameters for each condition. The heuristic oracle model assumes that people know the correct answer (thus "oracle"), but produce errors that vary by condition without regard to individual trial details ("heuristic"). These errors include some bias towards the center (given by a linear relationship between average reported position and the deterministic end point), and response variability distributed around that shifted position (with variance estimated independently for each condition). In other words, the heuristic oracle model is a non-physical error model. This model can capture the gross 'shift' in expected position that was observed in the data in each condition (see Figure 3.3), but does not treat the shift as an inference done independently on each trial. The spread in responses was assumed to be constant within each condition, and was set at the average empirical standard deviation from that condition. Like the stochastic models, this model was fit on three-quarters of the trials and tested on the remaining data.

Table 3.3 shows cross-validation likelihood for the four models. All log-likelihoods are shown as improvement over a baseline assuming that all data came from a single Gaussian. In addition, we included a 'perfect trial fit' model that knows the mean and standard deviation of responses for each trial – this serves as the plausible upper limit

Table 3.3: Model prediction of left-out data

Model	ΔLLH
Full	2,588
Dynamic	2,568
Perceptual	2,197
Heuristic Oracle	2,326
Perfect Trial Fit	3,259

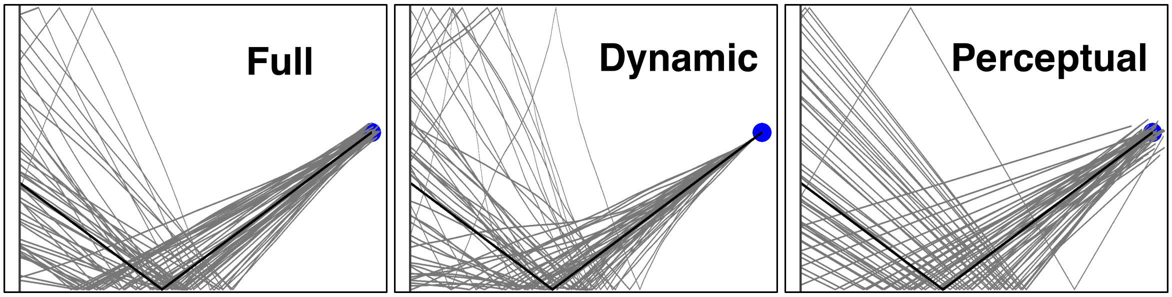


Figure 3.4: Sample simulation paths for one trial with each model. The grey lines represent individual simulations, the black line represents deterministic simulation. There is no initial uncertainty in the dynamic model, but it builds quickly over time, resulting in wavy paths. The initial position and velocity vary significantly in the perceptual model, but once started, the simulation unfolds deterministically. The full model uses both types of uncertainty and so has more certainty in starting positions than the perceptual model and straighter paths than the dynamic model.

on how well different models might do. The full stochastic model does best, followed closely by a model including only dynamic noise. Both the perceptual noise model and the non-physical model perform worse by many orders of magnitude.

The dynamic model performed nearly as well as the full model for two reasons. First, the parameter representing error in the initial position (σ_p) was set to a small value in the full model and explained very little of the variance in simulations. Second, much of the noise in initial velocity direction (κ_v) can be captured by increasing dynamic velocity noise (κ_m), and so we cannot say whether any initial velocity noise is required. The model with only perceptual noise did quite poorly because subjects' performance

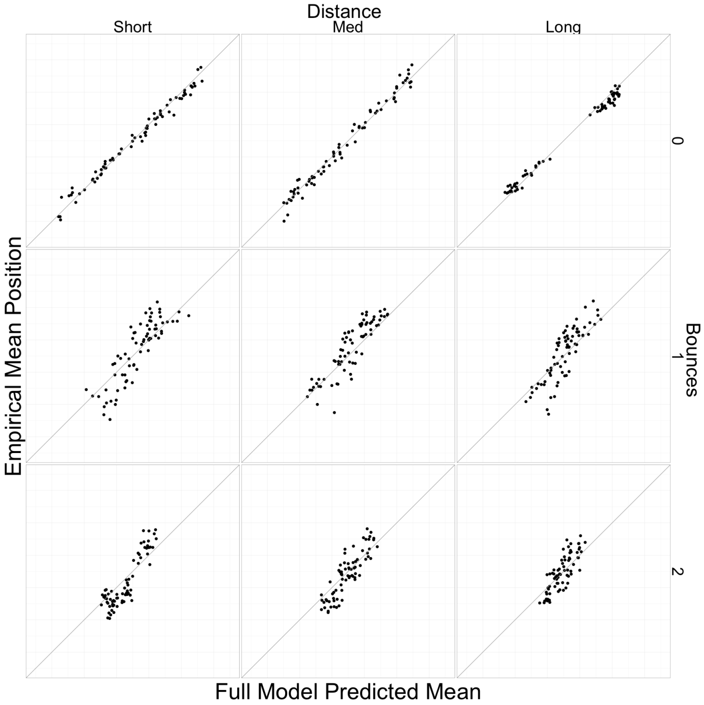


Figure 3.5: Full model predictions vs. empirical mean position by condition. Each point is a separate trial. The full model captures peoples' empirical behavior (including the center bias) well over every trial condition.

changed with each additional bounce, and thus human performance cannot be captured without dynamic uncertainty.

Trial-level simulations

Human predictions about individual trials within the same distance-by-bounce condition varied significantly: some had much larger variations in responses or greater shifts toward the center than others. These differences arose from trajectory characteristics other than total distance traveled or number of bounces. For instance, it is harder to predict the end position of a ball that bounces in a corner or balls that approach the paddle at a steep angle. If the stochastic physics model is capturing characteristics of intuitive physics, then it should capture this within-condition variability as well.

Table 3.4: Correlation between model and empirical by-trial means within condition

		Full			Heuristic Oracle		
		Distance			Distance		
		600	800	1000	600	800	1000
Bounces	0	.99	.99	.99	.99	.99	.99
	1	.86	.88	.85	.88	.77	.68
	2	.89	.87	.82	.82	.68	.45

The full stochastic model fit the variation in mean paddle position across trials well ($r = 0.93$), and slightly better than the predictions of the heuristic oracle model ($r = 0.90$). However, the difference between models is highlighted when considering individual conditions: although both models account for the mean position in the no-bounce conditions, only the full model continues to perform well as bounces and distance are added (see Table 3.4).

The standard deviation of predictions from the full stochastic model was well correlated with the standard deviation of subjects' responses across trials ($r = 0.79$, see Figure 3.6), albeit with a tendency to overestimate. Moreover, the stochastic physical model also captures the variability across trials within each distance-by-bounce condition (Table 3.5). Together, these results indicate that human uncertainty about final outcomes accumulates in a manner qualitatively similar to that predicted by a stochastic physical model.

In the experimental data, the amount of mean-shifting for each trial is related to the variance of the observations from that trial (Spearman's rho = 0.30), suggesting that people hedge their guesses towards the middle more as the amount of uncertainty increases. A center-prior captures this behavior by causing more reliance on the prior when there is a wider distribution of model simulations. This has the effect of shifting guesses more towards the center when physical simulations are more uncertain. The

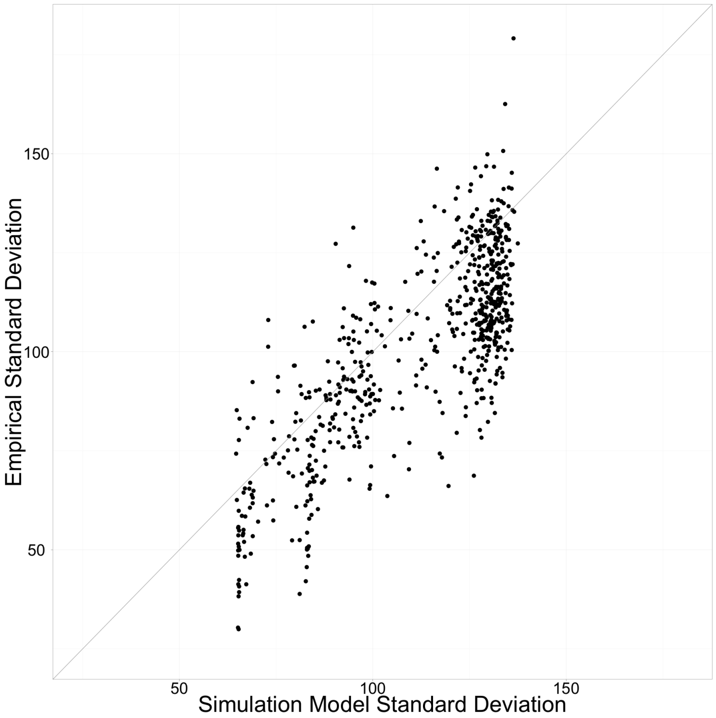


Figure 3.6: Full model vs. empirical standard deviation by trial. Each point represents a separate trial.

Table 3.5: Correlation between full model and empirical by-trial standard deviations within condition

		Distance		
		600	800	1000
Bounces	0	.54	.43	.17
	1	.53	.44	.30
	2	.14	.16	.17

stochastic physics model captures this phenomenon by predicting trial-level differences in uncertainty, and is thus better able to describe variation in human responses across trials than a constant mean-shift for each condition (see Figure 3.5).

3.4.3 Source of the center bias

Subjects positioned their paddle closer to the middle of the screen than where the ball actually ended, and we suggest that this bias arises from subjects' prior expectation that the ball will end in the center. In this section we address alternate explanations for this bias: is the center bias arising from task demands and strategies for dealing with this difficult task? Or is such a bias learned over time? We argue that neither of these accounts explains the bias we observe.

We assume that the middle of the paddle is each subject's best guess for the end position of the ball, but subjects could instead be attempting to minimize a loss function on the distance between each of the simulation outputs and where they place the paddle. Because predicted end-points under a physical model are somewhat skewed away from the edges (toward the center) due to the physical non-linearities of bounces, estimated positions will also be skewed toward the center relative to the modes of the distributions. However, these effects do not explain subjects' center shift. With a quadratic loss function (L2), the best placement of the paddle would be the mean of the simulations (Strook,

Table 3.6: Autocorrelation of paddle position with prior position

		Distance		
		600	800	1000
Bounces	0	.12	.14	.18
	1	.14	.17	.13
	2	.20	.20	.18

2011, p. 43), but as noted previously, the mean of the distributions was often centered on the end position of the ball and does not account for the observed center bias (indeed, this is why we suspected that subjects were using a center-prior). With a linear loss function (L1), the optimal paddle placement is the median of the simulation distribution, and a skew towards the center makes the median closer to the edges than the mean, predicting a relative edge bias. Although a more exotic loss function (e.g., L4 or L8) might increase predicted center-shifting, an arbitrary choice of this function would require more explanation than a center-prior.

Subjects may also have failed to move their paddle on some trials or not moved it quickly enough. Such a process would average out to yield an apparent center bias. If subjects' failure to move the paddle were exacerbated on more difficult trials, the center-shifting would be greater on those trials. We can test for such failures to move the paddle by assessing the autocorrelation between paddle positions on adjacent trials: on this account, the autocorrelation should be related to the amount of center-shifting. As can be seen in Table 3.6, this autocorrelation is low, although it does increase somewhat with the distance or number of bounces. However, it does not increase as center-shifting does – correlation with each condition's center-shifting (Table 3.1) is low and not statistically significant (Spearman's rho = 0.25; one-tailed permutation test, $p = 0.25$). Thus while movement failures may contribute somewhat to the center-shifting, they cannot fully explain it.

To make the next trial easier, subjects may have positioned their paddle closer to the center of the screen. This might make sense in a task where trials follow quickly after one another; and subjects have insufficient time to reposition the paddle between trials. However, we did not enforce inter-trial times in the experiment: subjects were free to move the paddle after each trial, and each trial was only started once the subject clicked the mouse. Furthermore, as evidenced by low autocorrelations between paddle positions, subjects do not appear to have any difficulty repositioning the paddle from one trial to the next. Thus it seems unlikely that such a strategy would benefit subjects.

Beyond this bias being imposed by task demands, it is possible that this expectation about the ball's movement was learned from the experiment. Each of the trials in the experiment was created with the constraint that the ball would not cross the plane of the paddle at the extreme ends of the screen; subjects may have noticed this fact and adjusted their responses appropriately. In order to address this concern, we tested whether the center-bias increased over the course of the experiment. We measured the amount of relative center-shifting that each subject had for each trial, and regressed this against the trial order, controlling for effects of the specific trial type; however, we found no evidence of a linear relationship between order and amount of center-shifting ($F(1,32996) = 0.139, p = 0.71$). Moreover, the estimated slope of this line suggests that, if anything, the center bias decreased over time.

Because the center-shifting behavior cannot be fully explained by task demands, and because this behavior did not change over the course of the experiment, we believe that the center bias is evidence of subjects' prior expectations about the ball's movement.

3.5 Discussion

We found that human performance on a physical prediction task is captured by a model of stochastic physics with a prior expectation about the final position of objects. Furthermore, we found that bias and variability of human predictions are driven by uncertainty about the dynamics: people use stochastic, rather than deterministic, physics to make predictions. This result supports recent findings that people predict object dynamics using unbiased intuitive physics models (e.g., Hamrick et al., 2011), and suggests two refinements to this view. First, the internal physics models themselves must be stochastic rather than rely solely on perceptual uncertainty to demonstrate non-determinism. Second, people do not directly use predictions from their physical models, but combine them with simple priors to produce rich behaviors.

Though we found that dynamic uncertainty contributes substantially to predictions in this task, we do not know how people might adjust this uncertainty based on task demands. In this experiment, the ball was easy to see (low perceptual uncertainty) and the background was uniform (suggesting less perturbation during movement). Lower contrast between object and background might cause greater perceptual uncertainty; likewise, backgrounds suggesting a rough surface might cause people to introduce more stochastic movement error into their simulations. An interesting direction for future work is to explore how people adjust the uncertainty within their intuitive physics models to account for different expectations about the world.

We also found that people modulate their physical predictions via prior expectations about the outcomes. Although these expectations could arise in many ways, here we were able to capture human behavior well by using a simple expectation about the final position: people believed that the ball was more likely to end up in the center of the screen. This expectation might arise because in similar games such as air hockey opponents are more likely to shoot the puck towards the goal in the center. However, it is also possible that this is an approximation of other sorts of priors (e.g., objects tend to travel in a more horizontal direction). More research is required to understand exactly what these prior expectations are, how they develop, and under what conditions they become integrated into models of intuitive physics. Regardless of the prior used, we think that this might reflect a more general strategy that people may adopt to account for their uncertainty in their internal physical model itself: by adjusting model predictions via a simple prior on outcomes, behavior will be more robust to errors in the simulation model. A similar process may suggest a means for combining model-based and model-free predictions (Glascher, Daw, Dayan, & O'Doherty, 2010): learning simple expectations about the world is a good hedge against model error.

Our models predicted systematically larger variances than those we observed. This may be due to our simplistic choice of the shape of the prior. Gaussian cue combination of the prior and simulated distributions produces dependence between variance and mean-shift: a greater mean-shift arises only from greater variance. Thus to best fit the predicted means, using a Gaussian prior required a biased variance estimate. Further work is required to understand the priors people actually hold (e.g., Stocker & Simoncelli, 2006) to refine the models that people use to simulate the world.

This work supports the hypothesis that intuitive physics models can be built upon a Newtonian framework. Moreover, these models are not deterministic, but incorporate sources of dynamic uncertainty. Furthermore, people do not trust these models entirely, but combine their predictions with simple expectations about the outcome itself. Though just a first step, this provides a framework for disentangling and understanding the various components of intuitive physics models.

3.5.1 Acknowledgments

This work was supported by BIAL Foundation grant to Edward Vul.

Chapter 3, in full, is a reprint of the material as it appears in *Topics in Cognitive Science* 5(1), 2013. Smith, Kevin A; Vul, Edward. The thesis author was the primary investigator and author of this material.

References

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Glascher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585–95.

Hamrick, J., Battaglia, P., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2009). A Bayesian framework for modeling intuitive dynamics. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *31st annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

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

Strook, D. W. (2011). *Probability theory: An analytic view*. Cambridge: Cambridge University Press.

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

4 Accumulating physical evidence over time

Abstract

People regularly make and update their predictions about physical events in real-time – for instance, positioning oneself to return a tennis shot requires not just predicting where the ball will go, but also quickly updating that prediction in light of new evidence such as the ball nicking the top of the net. But the information processing required to keep track of these predictions and update them based on ongoing evidence is far from trivial; how then are people able to do this information processing in real-time? Here we investigate how people perform this ongoing physical prediction with a novel task that required participants to make and update their predictions about the future state of the world in real-time. We find that peoples' predictions are consistent with a theory that they accumulate evidence over time from an 'intuitive physics engine' that noisily simulates the future state of the world. This demonstrates that human interactions with the world can be supported by rich, structured predictions about what might occur that are continually updated as the world provides additional information.

4.1 Introduction

A good hockey player plays where the puck is. A great hockey player plays where the puck is going to be. – Wayne Gretzky

In hockey, the puck can move across the rink in a matter of seconds. A hockey player who only skated towards where she sees the puck would spend the entire game chasing it and rarely reaching it; instead, she must build predictions about where the puck will be so that she can skate there. But she must also update these predictions as the game unfolds: if she thinks the puck will cross the rink but she sees an opponent intercept it, she should form new predictions about where the puck will be and update her own trajectory accordingly. And crucially, all of these predictions must be done in real-time. Of course, this capacity is not limited to hockey players: when driving, we track and predict the motion of other cars and pedestrians, and even children playing tag must track and predict each others' behavior.

Yet generating and updating predictions about where the puck will be is a difficult task. It requires not just tracking the puck over time, but also extrapolating the trajectory of the puck through bounces off the sides of the rink or other players, and constant updating of that extrapolation to integrate new information. And these predictions must be formed in real-time while the world itself is changing. How then do people perform this challenging task?

Prior research suggests that people track objects using a limited form of prediction: integrating noisy estimates of the position and trajectory of objects to probabilistically determine where the objects will be observed next, then updating beliefs about the position and trajectory in light of those future observations (e.g, Kwon, Tadin, & Knill, 2015; Vul, Frank, Alvarez, & Tenenbaum, 2009). Because our beliefs about the new location of an object are a weighted combination of our prior expectations and observations, our tracking is stabilized and not unduly reliant on potentially noisy observations. However, the predictions required for object tracking tasks are very short, since the object itself remains visible, and often do not require accounting for object interactions (e.g., the hockey puck bouncing off of the side of the rink; though for an example of tracking through bounces c.f. Hayhoe, Mennie, Sullivan, & Gorgos, 2005), so it is not clear how generalizable this tracking process is to predictions more than a few moments out.

On the other hand, recent studies of physical reasoning have suggested that people make physical predictions by taking noisy samples of the attributes of objects and simulating how the world will unfold with an ‘approximate physics engine’ (Battaglia, Hamrick, & Tenenbaum, 2013) – often termed the ‘noisy Newton’ hypothesis (Sanborn, Mansinghka, & Griffiths, 2013). But this theory is specified at Marr’s computational level (Marr, 1982), and therefore while it can explain what decisions we make when we reason about physical events – for example, how people judge causality (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012) or determine the positioning of objects (Scarfe & Glennerster, 2014) – it does not specify how those predictions should evolve over time as the world (and therefore our simulations) change.

Investigating how our physical predictions change therefore requires studying the dynamics of how we integrate evidence – not just *what* we predict will happen, but *when* we make those predictions. Hamrick, Smith, Griffiths, and Vul (2015) proposed that we sequentially produce simulations from our internal physics engines until the net evidence from those simulations in favor of one possible future is large enough to note it as a prediction. This model explained peoples’ decision times across a variety of conditions in a simple dichotomous prediction task (“if this ball continues its trajectory, will it pass through this hole in a wall, or bounce off?”), but this decision was made after viewing the relevant trajectory, and therefore cannot explain how this evidence dynamically evolves along with the world.

We therefore investigate here how people use simulations to make and update their predictions in a changing environment, and therefore how our evidence about what might occur evolves over time. Participants performed a task in which they judged the future behavior of a ball bouncing around a computerized table, and were asked to continually indicate their predictions as the ball moved around. We propose that people will accumulate evidence from their physical simulations similarly to Hamrick et al. (2015), but because the state of the world changes over time, so will the starting point of the simulations (and therefore the evidence they provide). We find that a model instantiating these principles predicted participants' judgments extremely well, and requires accounting for both physical simulation and accumulating evidence from those simulations to explain how people perform this task.

These findings provide insight into how people develop and update beliefs about the future when the world itself is changing: we use structured models of the world to make predictions about the future, and update our beliefs in line with changes in those predictions.

4.2 Experiment

4.2.1 Participants

One hundred participants were recruited from Amazon Mechanical Turk in accordance with UCSD IRB regulations. The recruiting and data recording was supported by the psiTurk framework (Coenen, Markant, Martin, & McDonnell, 2013). Participants were limited to those who had a US IP address, a minimum approval rating of 95%, and a non-mobile browser – the default exclusion settings for the psiTurk framework.

Participants were compensated $1 for their participation, which lasted approximately 15 minutes. Two participants were eliminated due to data recording errors, leaving data from ninety eight participants.

4.2.2 Methods

On each trial, participants saw a ball moving around a ‘table’ on the computer screen that contained blocks and both a red and a green target. The ball bounced perfectly elastically off of the edge of the table and blocks according to Newtonian mechanics (instantiated with the Chipmunk 2D physics engine; Lembcke, 2011), ending when the ball reached one of the two targets. While the trial progressed, participants were asked to predict whether the ball would hit the red target or the green target first, indicating their guess by holding down either the ‘z’ or the ‘m’ key (each key counterbalanced for red and green between participants). If they were unsure, participants could press neither key, and if their prediction changed mid-trial, they were encouraged to switch keys. Holding down a key would fill a bar of the associated color, and at the end of the trial, the score would be determined by the difference between the proportion of time the keys for each target were held down (with an offset to encourage participants by providing them with mostly positive scores):

$$Score = 20 + 100 * (Prop_{Correct} - Prop_{Incorrect}) \tag{4.1}$$

After each trial participants were notified of their score and could continue to the next trial by pressing the spacebar.

The tables were each 1000px wide and 620px tall to ensure that both the tables and controls would fit on screens with a minimum of 1024x768px resolution (which includes all resolutions of XGA and HD or better). If someone attempted to participate but did not have a screen of at least these dimensions, they would be notified that their screen was too small and were not allowed to continue.

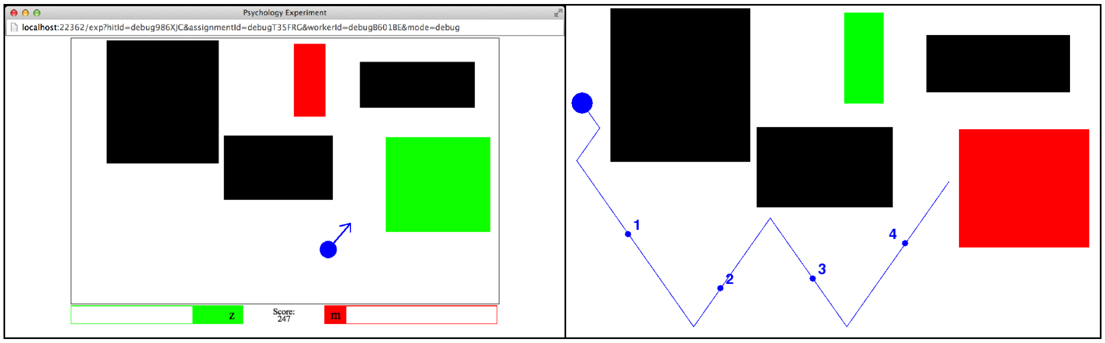

Psychology Experiment
localhost:22362/exp?hitId=debug986XJC&assignmentId=debugT35FRG&workerId=debugB601BE&mode=debug
z
Score:
247
m
1
2
3
4

Figure 4.1: *Left:* Screenshot of a trial. Participants viewed the ball moving around the screen and would hold down either ‘z’ or ‘m’ to fill the bar corresponding to the target they believed the ball would eventually reach. (The arrow was not visible to participants). *Right:* The path that the ball traveled on the example trial. The points and numbers indicate the time in seconds that the ball took to reach that point.

On all trials, the ball traveled at 300px per second. The display was updated every 25ms, but responses were polled only once every 100ms.

To ensure that online participants were paying attention through the experiment they were told to keep the window up at all times, and if their browser window lost focus (indicating that they were attending to another task) or they shrunk the window to below 1010x755px resolution they were reminded that the window should remain large and in front, and that trial was marked as invalid for data analysis. This process excluded 143 of the 9,800 trials.

4.2.3 Materials

All participants were given the same 100 trials, but with the order randomized. Of these trials, 95 were randomly generated subject to the constraint that they last between 2 and 10 seconds, and so that there were 19 trials each with one through five walls to ensure a range of difficulties and paths.

Of the remaining five trials, four were ‘contained’ trials in which the ball physically could not reach one of the two targets.[1] These trials were created to determine whether simulation-based physics could account for participants' judgments when reasoning about containment could provide an alternative solution without simulation.

All target colors were randomly swapped for each trial to avoid color bias effects. Responses in swapped trials were re-swapped so that analysis of predictions was consistent with goal locations rather than color.

4.2.4 Aggregate predictions

We analyzed participants' aggregate performance across trials via their total score (Equation 4.1). Participants mostly scored in the same range, with low variability in the average score each participant earned across all trials (mean = 56.9, sd =6.4, range = $[30.4, 70.0]$). Furthermore, participants' scores for each trial were very consistent (split half correlation, $r = 0.98$).[2] Because of this consistency we found it appropriate to explain how people make predictions on average rather than focus on individual differences.

On the other hand, the average score for each trial across all participants was highly variable (mean = 56.9, sd = 23.8, range = $[-25.6, 87.7]$), suggesting that we captured a range of difficulties, from trials where people were more likely to predict the incorrect target to trials where most people very quickly determined which target was correct.

However, this does not explain why participants make errors on some trials and not on others. To explain human predictions, we turn to a computational model of the

[1] Due to an error in the production one of the hand-crafted trials – the ball could in fact reach the 'inaccessible' goal – it was not treated as 'contained.'

[2] This statistic was calculated by splitting participants into two equal sized groups, determining the correlation of average trial scores between the groups, and averaging this correlation across a large number of possible splits. Thus it is a measure of how well we can predict participants' behavior on each trial from a separate set of participants.

prediction process.

4.3 Physical evidence accumulation

The model we used to predict behavior on this task has two parts: the physical simulator, which provides possible paths that the ball can take, and the evidence accumulator, which uses the output of the physical simulations to determine whether there is enough evidence in favor of one of the targets to make a decision.

4.3.1 Physical simulator

The part of the model that simulates the trajectory of the ball is based in large part on the model of Chapter 3 (Smith & Vul, 2013). This model assumes that people base their physical simulations on real accurate, Newtonian mechanics but must incorporate uncertainty about the world into their physical judgments.

This model captures two sources of uncertainty: 1) *perceptual uncertainty* arises from the noisiness of inferring the position and movement of objects, and 2) *dynamic uncertainty* is uncertainty about the roughness and elastic properties of the table and walls that could cause the ball's path to deviate from idealized Newtonian physics over time. See Chapter 3 for further details on the physical simulation and uncertainty parameters.

The physical simulator produces 100 simulation paths every tenth of a second for each trial to replicate the polling frequency in the experiment. Each simulation path would continue until the ball reached one of the two targets.[3] Therefore the physical simulator provided a proportion of the number of paths that reached the red versus green target at each time-step of each trial. This information was provided to the evidence

[3] For computational efficiency, if a simulation path lasted more than 60 seconds, it would end and be replaced by another path that reached one of the two targets within the limit.

accumulator to determine how quickly evidence for one target or another changed over time.

4.3.2 Evidence accumulator

While the physical simulator provides the probability of sampling a path that reached either of the targets at each time point, the evidence accumulator probabilistically samples from this distribution to build evidence for one target versus the other.

To model this decision process, we use a well-known decision policy for sample-based choices: the *sequential probability ratio test*, or SPRT (Wald, 1947). According to the SPRT, to make a dichotomous decision, an agent probabilistically samples evidence in favor of one hypothesis or the other, then makes a decision when the net evidence in favor of one hypothesis reaches a set threshold. SPRT (often called the *drift diffusion model*) is widely used to model the time-course of choices in simple decision making tasks (e.g., Ratcliff & Rouder, 1998; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006), but, crucially for this task, SPRT has been shown to explain the time course of physical decision making (Hamrick et al., 2015) and related models have been extended to non-stationary evidence (Tsetsos, Usher, & McClelland, 2011).

However, because the state of the world evolves over time, not all sampled evidence should be considered equal – a sample of where the ball will go from three seconds ago should be less informative than a current sample, since the current sample incorporates knowledge about the ball's current location and trajectory. We therefore assume that information *leaks* over time (Usher & McClelland, 2001), providing primacy for current evidence.

Therefore, this decision process can be formalized by assuming that at each time point t a sample path is drawn from the physical simulator, and the sample evidence (S_t) is set to 1 if that path reaches the red target, or -1 if that path reaches the green target.

The net evidence in favor of the red target (E_t) is updated based on that sample and the amount of leakage (L):

$$
\begin{aligned}
&E_t = E_{t-1} * (1 - L) + S_t \\
&E_0 = 0
\end{aligned}
\tag{4.2}
$$

The choice of which decision to make (D_t) is therefore given by whether the magnitude of the evidence has reached some threshold T:

$$
D_t = \begin{cases} \text{“Red”} & \text{if } E_t \geq T \\ \text{“Green”} & \text{if } E_t \leq -T \\ \text{“None”} & \text{otherwise} \end{cases}
\tag{4.3}
$$

Because there is no analytic solution for the probability of making a decision, this evidence accumulation process was repeated 1,000 times for each trial to determine the probability of making each decision at each time point.

Finally, because people cannot instantaneously process the information on the computer screen, form simulations, and push the button (and indeed the timing of each of these elements will differ from person to person), we accounted for all of this time unrelated to evidence accumulation by offsetting all decision probabilities by constant amount t_{off}, then convolving those predictions with a Gaussian kernel with standard deviation t_{width} to account for variability in integration and motor timing.

4.3.3 Explaining participants predictions with physical simulation

Participants’ predictions at each time-step of each trial were well explained by accumulating evidence from a physical simulation engine: the model could explain both

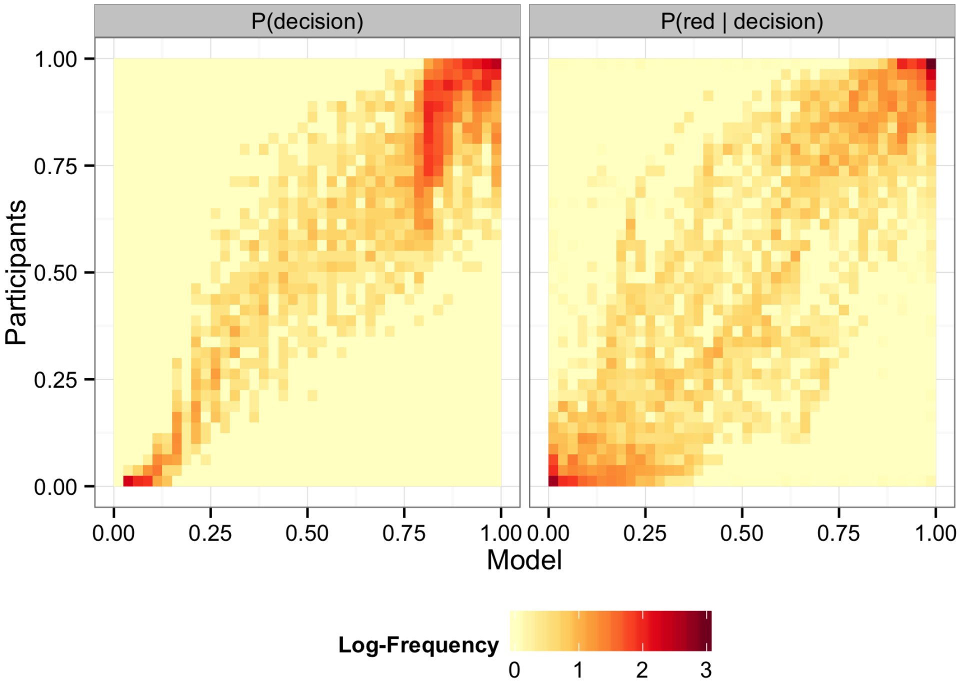


Figure 4.2: Joint histogram of model (x-axis) and human (y-axis) decisions. *Left:* the probability of making any guess (pushing a button). *Right:* the probability of choosing 'red' given a decision. Colors indicate the log-frequency of time points in each bucket (weighted by $P(decision)$ for $P(red|decision)$ to avoid overweighting buckets based on a small number of color decisions), with redder colors indicating more observations. Observations along the diagonal indicate the model is accurately capturing the exact proportions of participants making those judgements.

how often participants decided to push either button (indicating they had confidence in a prediction, $r = 0.950$, 95% CI = $[0.946, 0.953]$), and if they made a decision, which goal they believed the ball would go into (weighted $r = 0.952$, 95% CI = $[0.949, 0.955]$; see Figure 4.2).[4]

This also explains why some trials are more difficult than others – on some trials peoples' physical simulations almost all headed towards a single target, so evidence

[4] Many time steps had very few participants indicating any decision, and therefore estimates of the empirical probability of deciding 'red' versus 'green' in those cases would necessarily be imprecise. To adjust for the imprecision in these estimate, we calculated the correlation between $P(red|decision)$ between participants and the model weighted by the number of observed decisions for that time step (Pozzi, Di Matteo, & Aste, 2012).

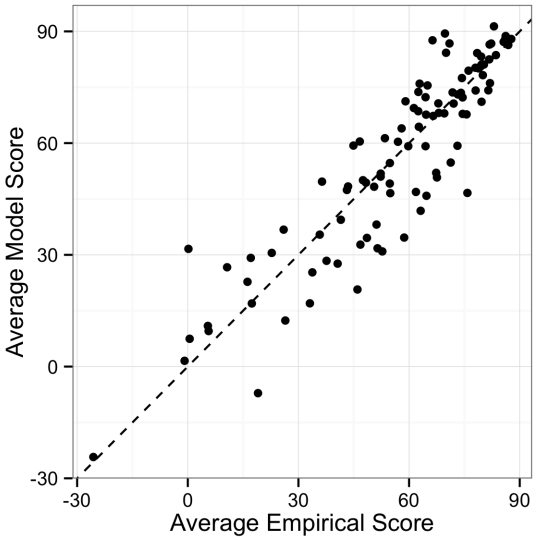


Figure 4.3: Participants' average scores on a trial versus predicted scores from the physical simulation model. Each point represents a single trial. The model was unbiased and well correlated with participants' scores ($r = 0.90$).

would build quickly for that target, leading to a high score, while on other trials the simulations were more uncertain or might even lead people to believe the ball would go to the incorrect target. Therefore, the average score earned on each trial by all of the participants was well correlated with the average score expected if they were accumulating evidence from simulation ($r = 0.90$, 95% CI=$[0.853, 0.934]$) and unbiased (average participant score = 56.9, average model score = 56.2; see Figure 4.3).

In addition, the physical simulation model captured the qualitative dynamics of many of the trials. On some trials, participants believed that the ball would head to one target before they switched their predictions to the other target, and in some cases switched their predictions back again. These decision switch points often occurred at junctures when 'surprising' events occurred – for instance, when the ball hit the corner of a wall and bounced at an odd angle (e.g., Figure 4.4, top), or narrowly missed a wall (e.g., Figure 4.4, second trial from top). Noisy physical simulation explains why people find these events to be surprising: stochastic ball paths will often hit the wall at a point that is not the corner, or will hit the wall instead of narrowly missing it, and therefore after these events occur, internal physical simulations provide radically different evidence than they had previously. Therefore, the physical simulation model's predictions follow both the time-course and magnitude of participants' predictions, explaining why some trials are more difficult for people.

4.4 Alternative decision heuristics

In the preceding section we demonstrated that ongoing human physical prediction can be well explained using a model of accumulating evidence from a noisy physical simulator, which suggests that this mechanism underlies physical reasoning. However, to use this as evidence of the underlying mental processes, we must show not just that it explains physical prediction well, but that it explains human predictions better than alternative models. We therefore propose two alternative mechanisms by which people might make their physical predictions: (1) people still use noisy physical simulations but base predictions only on instantaneous evidence, and (2) people solely use surface-level features of the trials to make predictions without relying on physical simulation at all. These comparisons demonstrate that physical prediction requires accumulating evidence over time, and that it relies on simulation, respectively.[5]

[5] A third alternative that was considered but not modeled was that people keep all prior evidence without discounting. This assumption was obviously wrong as it could not explain how people switch predictions quickly midway through a trial – if people had simulated three seconds of the ball going towards the red goal then simulations switched and began ending on the green goal, it took them significantly less than three seconds to switch their decisions, suggesting that not all evidence was weighted equally.

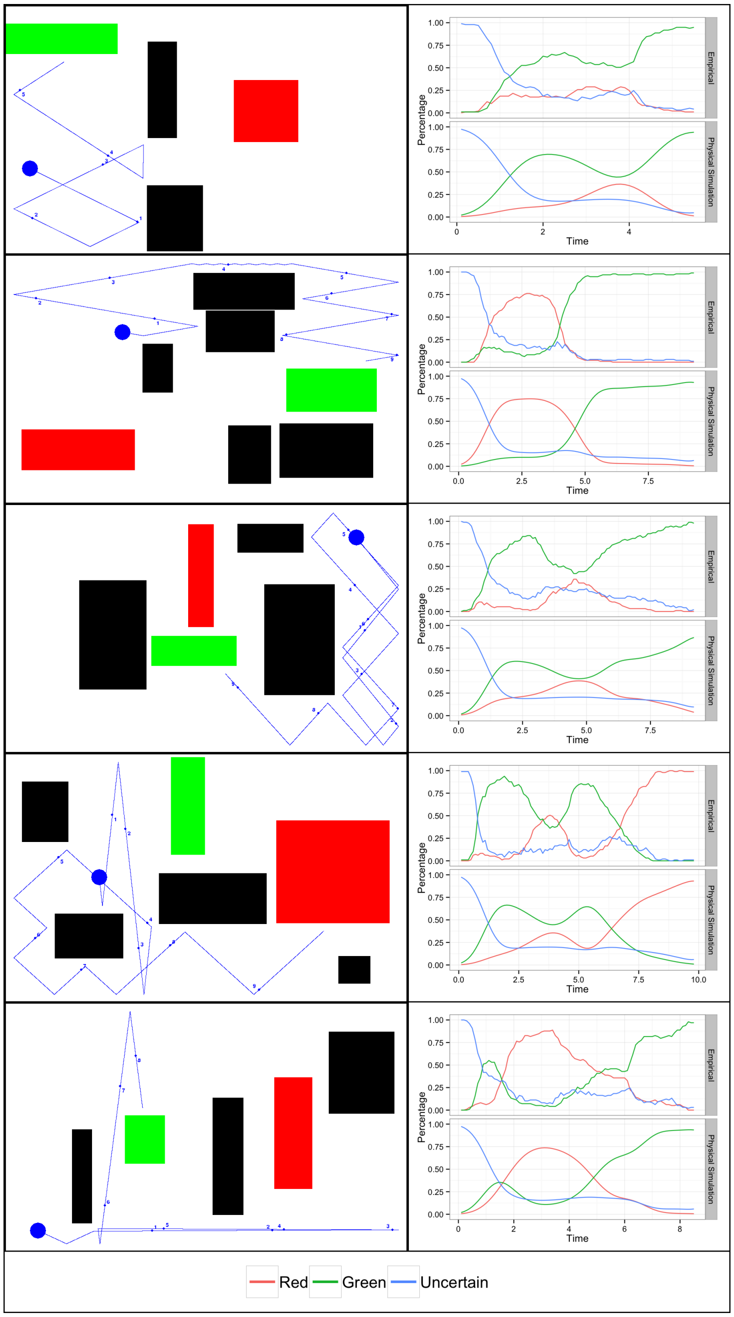


Figure 4.4: Sample trials with changes in confidence. *Left:* the path that the ball traveled during that trial, with the numbers representing the time in seconds that the ball would pass each point. *Right:* the proportion of people pushing either or no button (*top*) or the model's distribution of belief about how many people should be pushing each button (*bottom*) over time. Accumulating uncertain simulation evidence can explain changes in beliefs well; often these 'surprising' moments occur when the ball narrowly misses or hits a wall, or takes a bounce off of the corner of the wall.

4.4.1 Instantaneous physical evidence model

We have suggested that peoples' predictions are based on accumulating evidence over time. While integrating older information into our predictions can be a useful way to stabilize those predictions so that they do not vary wildly with new, conflicting information, it can also cause slower reactions to that new information if the world has in fact changed. It is therefore important to consider whether we might use only the most recent simulation information, rather than partially rely on prior expectations.

This alternative is instantiated in the 'instantaneous physical evidence model.' similar to the evidence accumulation model, the instantaneous physical evidence model consists of a physical simulator and a mechanism for making decisions. Here we assume that people use a similar SPRT decision making mechanism based on the outcome of their simulations, but that instead of integrating information over time, they take a number of samples from only the most recent simulation point. However, they use the same decision rule as the evidence accumulation model: deciding 'red' or 'green' if the net evidence exceeds a threshold, and abstaining from a decision if it does not. Similar to the model with evidence accumulation, we applied a time offset and smoothing to account for variable time to set up simulations and make a motor response.

4.4.2 Surface-level feature model

While a large set of studies suggest that we use simulation to make predictions about physical events, there is some evidence that we do so with sparse evidence – for instance, Hamrick et al. (2015) found that we only need a net evidence of two simulations in favor of an event to be confident enough to predict that event will occur. But the evidence accumulation model suggests that we are continuously producing new simulations. This might therefore be an overly taxing cognitive task, and so we consider

Table 4.1: Predictors used in the surface-level feature model

Predictor Type	**Predictor**
Trial features	Log-area of the red goal
	Log-area of the green goal
	Log-area of the walls
	Proportion of the screen that is clear of any walls/goals
	The shortest (non-physical) path from the ball start to the nearest goal
	Whether the shortest path reaches the red or green goal
	Whether it is a contained trial
Current trial state	Average distance to the goals
	Difference in distance to the goals
	Offset of velocity vector from the midpoint between the goals
	Angular difference between the velocity vector and goals
	Smallest difference in velocity vector angle from a cardinal direction
	Total number of walls blocking the goals (in a straight line)
	Difference in the number of walls blocking each goal
Prior occurrences	Log-time since the start of the trial
	Number of bounces the ball has taken so far

whether people might use surface-level heuristics without any simulation to accomplish this ongoing prediction.

The surface-level feature model was designed as a multinomial logistic regression to decide whether to predict ‘red’, predict ‘green’, or abstain from pushing any button. Because there have been no prior proposals of specific heuristics that people might be using in this task, we included sixteen predictors in this regression that captured a wide range of potential surface-level features that we believed might impact predictions (based on general features of that trial, the current state of the trial, and past occurrences; see Table 4.1). Finally, we applied the same time offset and smoothing as we had for the other models.

4.4.3 Model comparisons

The physical evidence accumulation explained participants' behavior better than both the surface-level heuristics model ($\Delta LLH = 9,797$) and the instantaneous physical evidence model ($\Delta LLH = 20,943$).[6]

Only the physical evidence accumulation model could explain both when people would make a decision to push a button ($r = 0.950$) and which target they believed the ball would go into (weighted $r = 0.952$; see Section 4.3.3). The surface-level heuristics model could explain participants' decision to make a prediction slightly better ($r = 0.963$, 95% CI=$[0.961, 0.965]$), but did not explain as well which target participants believed the ball would reach (weighted $r = 0.890$, 95% CI=$[0.883, 0.897]$; see Figure 4.5, bottom). Furthermore, many of the predictors from the surface-level model were correlated with the evidence accumulation model. For instance, strong predictors for pushing the red button in the surface-level model were (a) the ball was closer to the red target than green, (b) it was heading towards the red target but not the green and (c) there were no walls between the ball and the red target but were walls blocking green – exactly the conditions when most simulation paths would be expected to reach the red target.[7]

Conversely, the instantaneous physics model could not explain when participants would push a button as well ($r = 0.931$, 95% CI=$[0.927, 0.935]$), but performed equally well at determining which goal they believed the ball would reach (weighted $r = 0.952$, 95% CI=$[0.949, 0.956]$; see Figure 4.5, middle).

Together, these results suggest that noisy physical simulation is required to determine where the ball will go in the future, and that this evidence accumulates over time rather than being based only on the most recent simulation. The ability of the

[6] Because the evidence accumulation model had an equal number of parameters to the instantaneous physical model, and 24 fewer parameters than the surface-level model, any increase in log-likelihood would guarantee that the evidence accumulation model would have a lower AIC and BIC, suggesting that it is the most parsimonious model for the data.

[7] The converse predictions held true for selecting the green target.

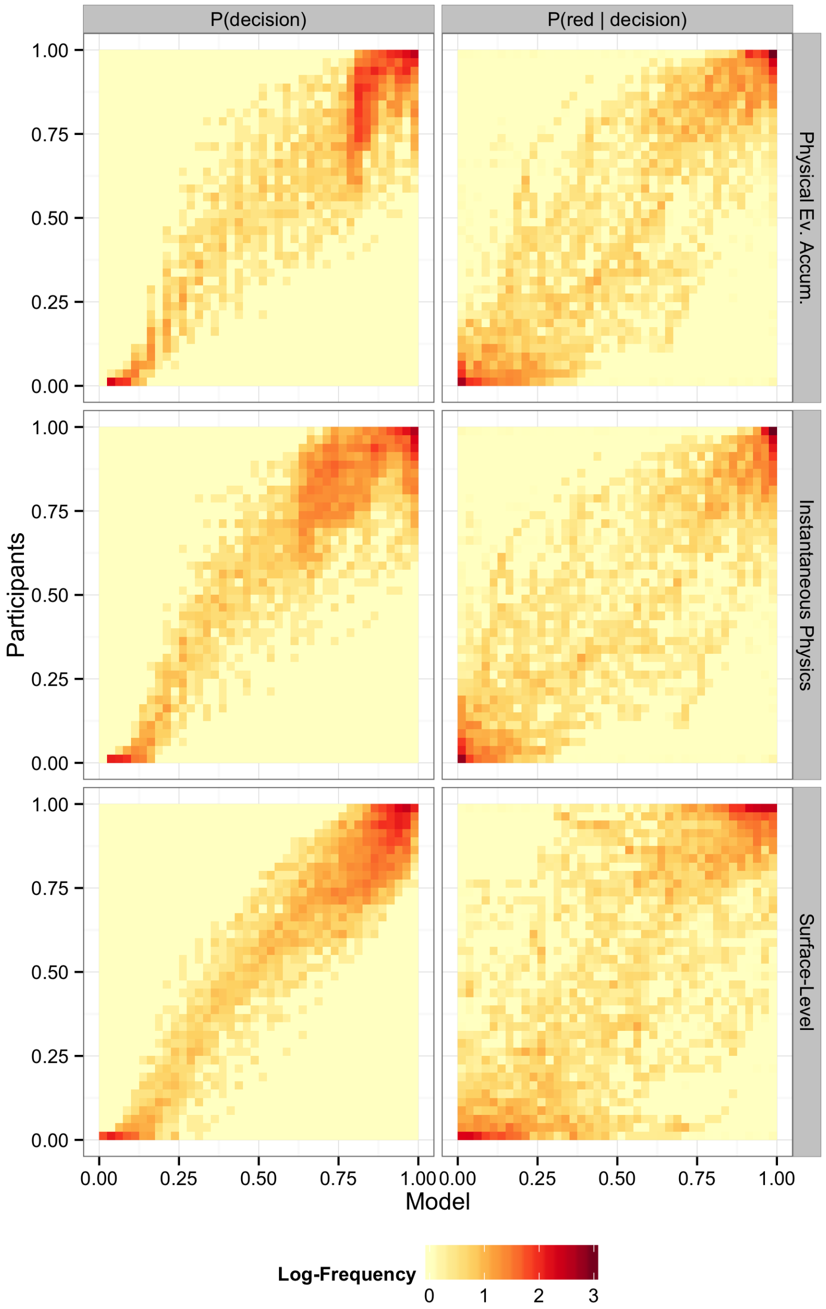


Figure 4.5: Joint histogram of model (x-axis) and human (y-axis) decisions for all models, calculated the same as Figure 4.2. *Left:* the probability of making any guess (pushing a button). *Right:* the probability of choosing 'red' given a decision. *Top:* the physical evidence accumulation model. *Middle:* the instantaneous physics model. *Bottom:* the surface-level heuristics model. Only the evidence accumulation model could explain both when participants made any guess and which target they guessed.

surface-level heuristics model to explain the propensity of participants to make any decision slightly better than the evidence accumulation model suggests that physical evidence accumulation might not be the sole determinant of participants' confidence in their simulations over all trials, but this simulation evidence accumulation process does explain the majority of how people decide that they are confident enough to indicate their predictions.

4.5 Discussion

Here we measured how people make and update physical predictions over time, and demonstrate that these predictions can be explained as people accumulating evidence from ongoing simulations of how the world will unfold. Furthermore, we show that noisy physical simulation is required to explain how people distribute their beliefs about what will happen in the future, and that people update evidence from these simulations over time rather than instantaneously change their predictions due to updated simulations.

However, while this study finds that ongoing human predictions are consistent with accumulating evidence from constant simulations, it is possible that the mind takes many fewer simulated samples to approximate this process. After all, in many cases a new simulation will provide roughly the same evidence as the preceding simulation, yet would take additional cognitive effort – constantly re-estimating the trajectory of a hockey puck slowly sliding towards the goal without interference, for instance, will likely do little to change your mind once you already believe that the puck will reach the goal. The mind might exploit these regularities to approximate consistent simulation with less effort, perhaps only producing new simulations after observing an event unexpected according to prior beliefs.

In addition, we looked for cases in which people might be making predictions by some mechanism other than simulation. We specifically seeded the experiment with four ‘contained’ trials in which the ball could not reach one of the two targets. Because this implies that all simulations will always indicate one of the targets, the evidence accumulation model should perform identically for all of these trials; however, participants’ responses did differ depending on the scene layout, with participants responding slightly more quickly than simulation would suggest on one trial, and much more slowly on another (see Figure 4.6). While this is a qualitative analysis based on limited data and therefore simply suggestive, it does raise the possibility that people make physical predictions through multiple routes, including reasoning about containment when it is easier than physical simulation (e.g. Davis, Marcus, & Chen, 2013). However, more work is required to explicitly tease apart when people might be using different types of reasoning.

The information processing required to perform real-time predictions is extremely challenging, yet people appear to do it with little effort. This work provides a step towards describing how this is accomplished: we use rich, structured models to determine how the world might unfold, and regularly update this information in light of changes to the world. Knowing how we keep in mind and update our predictions is necessary to understand how we can flexibly plan actions while also reacting to new information, and therefore how we so capably adapt to the world around us.

4.5.1 Acknowledgments

Chapter 4, in part, is currently being prepared for submission for publication of the material. Smith, Kevin A; Dechter, Eyal; Tenenbaum, Joshua B; Vul, Edward. The thesis author was the primary investigator and author of this material.

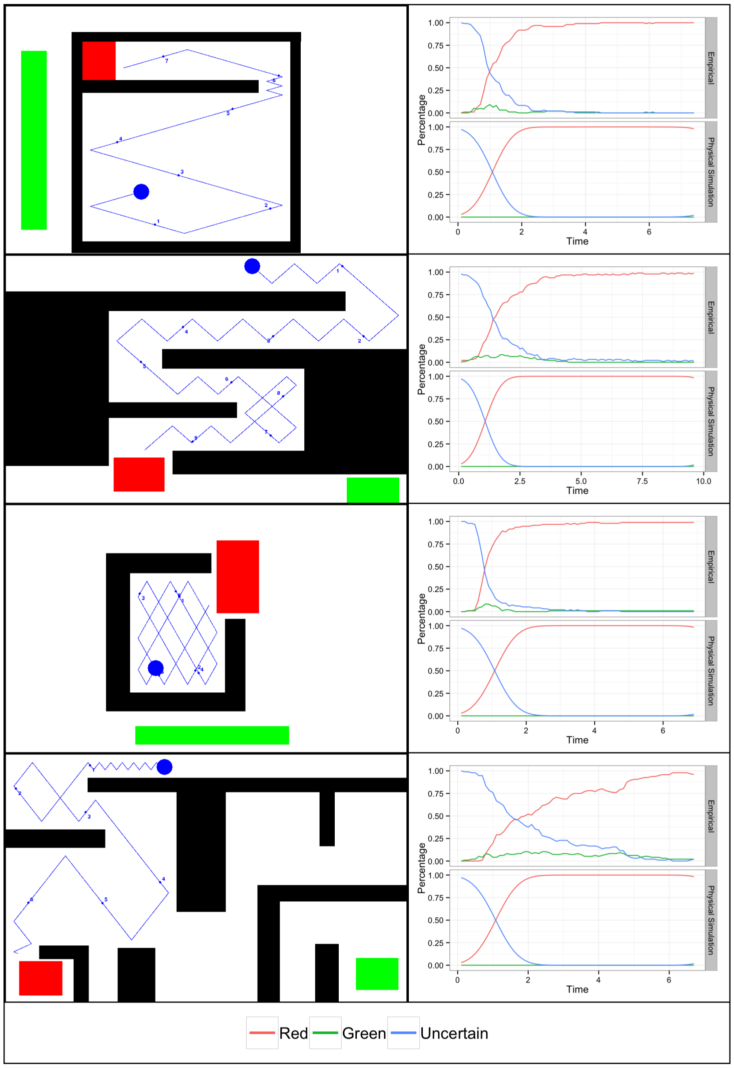


Figure 4.6: Contained trials. *Left:* the path that the ball traveled during that trial, with the numbers representing the time in seconds that the ball would pass each point. *Right:* the proportion of people pushing either or no button (*top*) or the model's distribution of belief about how many people should be pushing each button (*bottom*) over time. The model does not differentiate between these trials, since simulation paths will all end at the red target; however, the speed at which participants responded does change, with participants faster than suggested by simulation (e.g., second from bottom) or much slower (e.g., bottom).

References

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

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.

Coenen, A., Markant, D., Martin, J., & McDonnell, J. (2013). Using Mechanical Turk and psiTurk for dynamic web experiments. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 22–23). Austin, TX: Cognitive Science Society.

Davis, E., Marcus, G. F., & Chen, A. (2013). Reasoning from radically incomplete information: The case of containers. *Advances in Cognitive Systems*, *2*, 273–288.

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

Hamrick, J., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? Optimal mental simulation tracks problem difficulty. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hayhoe, M., Mennie, N., Sullivan, B., & Gorgos, K. (2005). The role of internal models and prediction in catching balls. In *Proceedings of the American Association for Artificial Intelligence.*

Kwon, O.-S., Tadin, D., & Knill, D. C. (2015). Unifying account of visual motion and position perception. *Proceedings of the National Academy of Sciences*, *112*(26), 8142–8147.

Marr, D. (1982). *Vision* [Book]. Cambridge, MA: MIT Press.

Pozzi, F., Di Matteo, T., & Aste, T. (2012). Exponential smoothing weighted correlations. *The European Physical Journal B*, *85*(6), 1-21.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

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

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

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

Tsetsos, K., Usher, M., & McClelland, J. L. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in Neuroscience*, *5*.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Vul, E., Frank, M. C., Alvarez, G. A., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22).

Wald, A. (1947). *Sequential Analysis*. New York, NY: Wiley.

5 General discussion

From a waiter stacking dishes only as high as they will balance, to a pedestrian crossing the street in front of a car, to a soccer player passing the ball to her teammate, peoples' interactions with the world are supported by expectations for how the world will unfold. I have argued that a core component of building these expectations is a rich simulation mechanism that extrapolates the future using approximately accurate physical principles, but must account for uncertainty about the properties of the world. In this thesis I studied three key facets of this reasoning process: when we use this simulation-based mechanism as opposed to other forms of reasoning, whether simulations are themselves stochastic, and how we accumulate information from simulation over time.

The prior literature on intuitive physics has been split, with some arguing that our physical reasoning is based on approximately accurate simulations, while others suggest that our core understanding of physical principles is erroneous. I resolved this disparity by studying when people use accurate versus erroneous principles. Studying peoples' concepts of a single physical principle (the trajectory of objects in ballistic motion) in three different ways, I found that when people are confronted with an interactive task, their predictions are based on accurate physical principles, but when they are asked to explicitly draw the ballistic trajectory, their predictions are erroneous and idiosyncratic. This suggests that we have multiple systems we use for physical reasoning, and the difference between the previously studied errors and our expert behavior is the underlying mental system we use to support them (Chapter 2).

While prior research suggests that our predictions are variable due to uncertainty in the world, there has been little study about where that uncertainty comes from: is it only initial uncertainty about the location and motion of objects that accumulates noise through deterministic simulations, or are our simulations themselves stochastic? By studying how peoples' predictions about where a computerized ball would go in response to changes in the initial motion and path, I teased apart the contribution of both initial perceptual uncertainty and dynamic uncertainty that accumulated in the process of simulation. I found that in addition to uncertainty about the visible properties of objects (e.g., their exact position and motion) people must account for uncertainty about unseen properties (e.g., the roughness of the floor may cause a ball to deviate from straight-line motion). This suggests that our simulations are themselves noisy, perhaps to account for unknowable deviations in the real world, such as imperfections in a ball causing it to take an odd bounce (Chapter 3).

Real-world physical reasoning does not occur at a single time, but rather unfolds along with changes in our environment. It is therefore important to understand how we use information from simulation over time to inform our predictions, and be able to explain the timing and evolution of our beliefs. I therefore measured how peoples' decisions about the future location of a computerized ball would change as they obtained further information about the ball's path. I found that the dynamics changes in ongoing predictions could be explained as consistent accumulation of evidence from their internal simulation engine, even as the outputs of the simulation engine change in response to updates to the world. Therefore, our beliefs about the future are a combination of our prior predictions and new information obtained from regularly produced simulations (Chapter 4).

While these findings have built a framework for understanding how people structure and use simulations to support physical reasoning, a full accounting of how we reason about the world around us requires tackling further challenges. For instance, determining how we decide to use simulation-based or other sorts of physical reasoning, or defining the algorithmic structure of how our physics engines run these simulations. Understanding these facets of physical reasoning can provide insight into many of the key questions of cognition, such as how the mind picks a strategy for solving a given problem, or how it accumulates evidence from complex prediction in an ever-changing environment.

5.1 Modes of physical reasoning

Although accumulating evidence from simulations can explain many of our predictions about physical events, this thesis and other work has shown that it clearly cannot explain the entirety of our reasoning. When we attempt to explain physical scenarios, our naïve explanations are often erroneous (e.g., Shanon, 1976; McCloskey, 1983), yet our interactions with our environment that rely on the same physical principles demonstrate a sophisticated understanding of the physics involved (Chapter 2). The dichotomy between erroneous and calibrated principles depending on the task suggests that we recruit different systems of reasoning to accomplish explanation versus interaction. When we can solve problems using scene parsing or reasoning, we may rely on those capabilities – we abstract rules to shortcut simulation (Schwartz & Black, 1996) or reason about containment rather than use simulation (Chapter 4, Davis, Marcus, & Chen, 2013) – again suggesting that there is more than one cognitive mechanism for physical reasoning. But why would we have more than one system to perform physical reasoning? Why might these systems rely on different physical principles? And how do we chose which system to use? These questions that grow out of the confluence of the many threads of research that study physical reasoning are an important next step to understand how we use all available mental tools to comprehend the world.

Differences in behavior when the principle is identical but the task is different are found in many domains beyond physical reasoning. For instance, when we explicitly make decisions about monetary bets, our judgments are typically not optimal according to economic theory, but if we make those same bets via a ‘motor lottery’ (being rewarded for touching a small target quickly or penalized for missing, so that the outcome is probabilistic), our behavior is much closer to optimal (Wu, Delgado, & Maloney, 2009). And this follows a well-known theory in the decision making literature: our fast, intuitive decisions are often thought to be based on a different system of reasoning than slower, deliberative choices (System 1 vs. System 2; Kahneman, 2011).

Often these dichotomies are characterized as ‘automatic systems’ (System 1) and ‘deliberative systems’ (System 2), and this split may be appropriate for physical reasoning as well. When we are throwing a ball to a friend we are unaware of the complex calculations that must be done to determine the exact force and angle that we will throw the ball with, yet when we solve simple high school physics problems many of us are well aware of the effort that it can take. Thus our ‘approximate physics engine’ may support automatic interactions with the world, while we need to use more deliberative reasoning to structure this knowledge into explanations.

5.1.1 Cognitive specialization of physical reasoning

While explicit reasoning about physics appears to be based on general purpose cognitive mechanisms, the approximate physics engine may be a specialized cognitive module, similar to language or face processing (Kanwisher, 2010). These specializations often perform highly complex tasks that we use often in our day-to-day interactions with the world, and therefore it is useful to have specific machinery that can perform this information processing quickly, though it is dedicated to a single task. Planning our actions around how the world will unfold requires split-second decisions – if we see a branch falling on us, we need to jump out of the way as quickly as possible – so attempting to reason about the future using general cognitive processes may take too long to allow us to survive in the world. Thus by having a specialization for physical reasoning, we are more easily able to plan our actions in a timely fashion.

Cognitive specializations typically have two common features. First, there are often regions of the brain that respond selectively when those modules are active – for instance, areas such as the temporo-parietal junction for theory of mind (Saxe & Kanwisher, 2003) or the fusiform face area for facial processing (Kanwisher, McDermott, & Chun, 1997). This would imply that a specialized brain area for physical simulation should exist, and indeed there is preliminary evidence that a set of brain regions in parietal and premotor cortex may underlie this type of physical reasoning (Fischer & Kanwisher, 2015).

Second, the modules typically are cognitively impenetrable – we cannot introspect into how they function. Just as we cannot explain how we conjugate verbs while speaking or what configuration of features makes something 'face-like', we cannot explain the rules that we use to update our simulations. In Chapter 2, I demonstrated that we at some level have access to accurate principles of ballistic motion, but when asked to explicitly note them through a drawing we display errors, suggesting it is not possible to directly query the workings of our simulation systems. This impenetrability might be due to the specialized nature of the processing – if these processes use specialized, efficient neural codes that cannot easily be interpreted by other high-level reasoning systems, or the underlying brain regions only send processed information to other areas of the brain, then introspection about the functioning of these regions should be impossible. Thus the specialized nature of physical processing might explain why we cannot reason about how we perform simulation.

While it is too early to claim that physical reasoning is a cognitive specialization (and it is still under debate whether *any* specialized cognitive systems exist – e.g., Huth, Nishimoto, Vu, & Gallant, 2012), this is a domain of knowledge that is a good candidate for specialization: physical reasoning underlies a wide variety of our interactions with the world (suggesting common usage), the simulation is cognitively impenetrable, and there is initial evidence that we have brain regions dedicated to physical reasoning.

5.1.2 Miscalibration of explicit knowledge of physics

Just because our explicit knowledge of physics utilizes a different system of understanding than physical simulations, it is not inherently obvious why our naïve theories should be erroneous. After all, if we base our theories on observations of the world, those observations are rooted in accurate physics. If we base our theories on what we imagine might occur in the world, then despite the cognitive impenetrability of the simulation system the imagined outcomes should also be accurate. Our errors therefore cannot come solely from our observations or imaginations, but rather must arise from biases in the process of interpreting these observations.

It is possible that these biases occur from generalizations about ourselves versus the external world. Many researchers claim that explicit physical reasoning is a process of analogy: solving problems in front of us by directly transferring the outcomes of similar situations we have observed in the past (Catrambone, Jones, Jonides, & Seifert, 1995), or learning general rules by abstracting between multiple related situations (Forbus & Gentner, 1986). But our own experiences are asymmetric, with our bodies providing us with privileged information about the forces involved when we touch or push an object. White (2012) suggests that this can explain certain 'impetus theory' beliefs: for instance, we often have to continually apply a force to an object to slide it along the floor, so we explain that as force ('impetus') causing velocity (instead of acceleration) and then believe that removing this force will cause motion to fade. Thus because we have some sense about how our interactions with the world 'feel' to us, we inappropriately use these observations to explain external motions.

5.1.3 Extracting accurate physical explanations from simulation

Although we cannot introspect about the rules that underlie physical simulation, we can still use this system to improve our explicit reasoning by imagining a physical event before making a verbal judgment about it. For instance, when people are asked about the relationship between pendulum length and period, they are typically inaccurate, but if they are asked to imagine the pendulum in motion first, their accuracy increases significantly (Frick, Huber, Reips, & Krist, 2005). Thus even though the rules of simulation are cognitively impenetrable, we can still access accurate mental imagery of physical scenarios that can be used in our explanations.

Schwartz and Black (1996) demonstrated that we can go beyond simple observations of our imaginations to abstract rules from simulations. They showed that when people are given word problems about gears (e.g., "If there are five gears lined up in a row and the leftmost gear turns clockwise, what will happen to the rightmost gear?"), initially they would use mental simulation to solve these problems, but given many similar problems they would eventually supplant these simulations with rules (e.g., "If there are an odd number of gears in a line, the leftmost and rightmost gears will turn in the same direction") and be able to respond significantly more quickly. Indeed, imagination plays an important part in modern techniques for teaching physics: introductory students are encouraged to learn basic principles by imagining how events that rely on those principles might play out, such as learning about inertia by imagining the feeling of stopping a large moving object (Helm, Gilbert, & Watts, 1985).

But if imagining physical events can correct for erroneous conceptions, then why do we continue to hold these incorrect explicit principles? Why do we not simply simulate an event before we provide an explanation for it? Understanding why we choose to use explicit reasoning versus physical simulation can help to resolve these questions.

5.1.4 The choice of physical reasoning systems

How we select which mental algorithm to use to solve arbitrary problems is a significant open question. Gershman, Horvitz, and Tenenbaum (2015) suggest the mind accomplishes this task by estimating the expected value of any given strategy (including the value of the correct answer, and the costs of waiting to provide that answer), then selecting the strategy with the highest expected value. However, this in turn requires a way of estimating the probability of getting a useful answer from a strategy before that strategy is applied, as well as how much effort would be required to produce that answer – a necessarily heuristic-based estimation.[1]

Whether this estimation tips in favor of the simulation system or explicit system is therefore very dependent on the goals of reasoning and the task environment. Depending on the heuristics for determining costs and benefits, this could lead us to produce incorrect answers in situations where accuracy is not critical. In an experimental study of naïve physics, participants who are asked to explain a physical principle are encouraged to convey an accurate picture of how the world works but there is no additional incentive for accuracy over an incorrect explanation – participants in Chapter 2, for instance, did not receive a monetary bonus for providing the correct drawing. Even if we know that producing an answer based on mental imagery of the event would be more likely to be accurate, if the costs of simulation outweigh the costs of determining the principle

[1] Although I argue that the principles of simulation are cognitively impenetrable, this does not imply that the expected costs and benefits are too. Even if we cannot introspect into how the system works, we can still take note of whether it provides us with useful information and how cognitively effortful it is.

explicitly, we may still choose the less accurate explanatory system if there is enough of a difference in effort.

But this theory assumes that simulation is more effortful than explanation, while intuitively simulation seems relatively effortless when we try to interact with the world. Perhaps a key difference is that it is effortful to set up a model of the world de novo: when we perform interactive tasks we already have information from our perceptual system that can initialize our simulations, whereas when we see a diagram or hear a verbal explanation we must construct a model of the world from scratch. This can explain why making judgments about the naturalness of an object's path is more accurate than picking out the very same path diagram (Kaiser, Proffitt, & Anderson, 1985; Kaiser, Proffitt, Whelan, & Hecht, 1992), since it is easier to simulate the path given the existing motion and compare the observed motion to those simulations than to set up a simulation without perceptual information and compare that path to a static diagram. However, given this theory we would also expect to see more accurate judgments in Chapter 2 when people drew trajectories after observing the system in motion; this might suggest that our heuristics for determining what system to use are in some cases suboptimal, or the enforced stop in the motion before asking participants to draw might have made simulation more difficult.

Therefore, an important line of research to understand physical reasoning is how these two systems trade off with one another. This thesis used simple dichotomies such as assuming interactive tasks would tap into simulation, while explicitly noting physical principles would tap into the explanatory system. However, the process by which people decide to use one system of physical reasoning over another is largely unknown. By studying how the mind assigns costs and benefits to each of these strategies across a variety of scenarios, we can discover not just how we decide to approach problems of physical reasoning, but can also shine light on the general process of 'meta-reasoning' about the cognitive strategies we use.

5.2 Approximations of physics

Throughout this thesis, I have argued that peoples' simulations are based on an 'approximate physics engine.' However, there are many different types of approximations, and to determine how we approximate physics requires answering many additional questions about the underlying cognitive processes. There are two types of approximation that need further study: approximations to the physics underlying the simulations, and approximations to probabilistic reasoning involving these simulations.

5.2.1 Approximations of physical principles

At its core, all of physics is an approximation to how the world works. Newtonian mechanics is very accurate for objects that we typically experience, but in the end is the result of many particles following the laws of quantum mechanics, and falls apart when objects are moving near the speed of light and relativity takes over. And many phenomena within the realm of Newtonian mechanics are simply the result of statistical regularities that can never perfectly describe the state of a system (e.g., fluid dynamics). Thus the question is not whether our internal models perfectly accurately capture the state of the world (they cannot), but rather *how* our internal models produce predictions that closely match the future.

However, there are multiple ways that a system might simulate approximately correct physical events. For an example of this we can look at the choices that go into computer physics engines, all of which are built to approximate physics in ways mostly undetectable to people. With only a simple rigid-body physics engine, there are a multitude of core choices to make (Millington, 2010): how much time should pass in a basic iteration between checking for collisions? Should simultaneous collisions be determined and resolved all at once, or sequentially? Are changes in motion calculated using forces or impulses? And these decisions only become more complex when we need to model soft-bodied objects (such as cloth) or fluids (Nealen, Müller, Keiser, Boxerman, & Carlson, 2006).

Fortunately, recent work has begun to investigate the structure of physical simulation. Collisions between objects, for instance, are very important. They are a core component of human physical conception (e.g., even newborns have a concept of *solidity* – that two objects cannot occupy the same space at the same time; Spelke, Breinlinger, Macomber, & Jacobson, 1992), and the choice of how to resolve collisions is one of the key decisions when designing a computer physics engine. Hamrick, Smith, Griffiths, and Vul (2015) have recently found that the process of determining how a ball will bounce off a wall (resolving that collision) takes additional time for people to process above and beyond extrapolating straight line motion. This finding is backed up by eye-tracking studies – Crespi, Robino, Silva, and de'Sperati (2012) found that if people view a short clip of the motion of a billiard ball and are asked to predict whether that ball would eventually pass over a specific spot on the table, people's eyes will follow the trajectory of the ball but will typically look more often at the sides of the table where the ball would bounce. Thus, just like with computer physics engines, resolving how two objects collide appears to require more processing from people than simple linear motion extrapolation.

Similarly, we must simulate fluids and be able to differentiate between types of liquid – for instance you might be more concerned if you spilled a cup of coffee near a stack of your papers than if you spilled a jar of honey, because the viscous honey might not flow far enough to ruin your work. While people can differentiate features of liquid based on low-level visual information (Kawabe, Maruya, Fleming, & Nishida, 2015), this would not explain how we might predict the flow of those liquids in the future. However, Bates, Yildirim, Tenenbaum, and Battaglia (2015) recently demonstrated that peoples' predictions about how different types of liquid would splash and flow through a set of obstacles could be well explained by a computer physics engine that uses a limited set of weakly connected 'particles' to simulate fluid dynamics, but not by shallow heuristics or deep learning mechanisms. This suggests a plausible algorithm that we might use to simulate the motion of fluids.

But while this recent work has begun to describe the structure of simulation, a large number of core questions remain. For instance, how well do our physical computations scale with complexity? How do we choose what parts of the world to incorporate into a specific simulation? And how fine-grained is the temporal resolution of our simulations? Answering these questions will bridge the gap from explaining *what* we accomplish with physical simulation to *how* we do it, and provide further answers into how the mind conceptualizes the world.

Complexity limits in simulation

Most research finds that our visual working memory is extremely limited – that we can only remember the specific properties of about three or four objects at a time (Luck & Vogel, 1997). Yet there are many cases where we reason about physical events that involve many more than four objects interacting. For instance, Battaglia, Hamrick, and Tenenbaum (2013) asked people to judge the stability of towers that could have ten or more independently moving blocks, and if we hope to simulate how a bucket of marbles dumped on the floor will behave, the number of objects simulated might reach a much greater number. But how is this possible if we can only account for four objects at a time?

One possibility is that we group objects into larger, connected objects based on heuristic assessments of how they might move. For instance, if we are simulating how a tower of blocks might fall, we might group a set of blocks together as the ‘base’ of the tower and treat that as a single object. Similarly, as we simulate marbles falling from a bucket, we might not treat them as separate entities, but rather assume that each marble is just part of a flow of objects. But how we decide to group objects in this way is still an open question.

However, a hint of how this might occur comes from more recent research into visual working memory. There is evidence that the low memory capacity estimates are due to the independence of objects enforced in these experiments – memory for color, for instance, is typically tested by asking people to remember a number of independently determined colors. But objects in the world typically are not as disjoint as these artificial stimuli, and people are much better at remembering objects if there is some degree of regularity to them (Brady, Konkle, & Alvarez, 2009). It is thought that we can do so by forming hierarchical representations of objects, and therefore remember ensemble representations (e.g., the average position and spread of a cluster of dots) rather than representing every item individually (Brady & Alvarez, 2011). Our ability to extract ensemble statistics extends beyond simple spatial information, allowing us to extract complex properties from sets of objects such as the average emotion in a crowd of faces (Haberman & Whitney, 2007).

Thus it is possible that this ensemble representation can explain complex physical simulation: to simulate the bucket of marbles, for instance, we cluster all of the marbles together based on similar properties and starting conditions, then simulate the motion of the ensemble rather than each individual object. The rules of this simulation would necessarily be different from the rules that govern simulating individual objects, as these would require simulating not just how the ensemble moves, but how it spreads out over time – especially as the marbles hit the floor. To perform this sort of simulation, we would need to extract ensemble information about physical properties, such as the average and variability of kinetic energy the marbles have at the point of collision, and use physical principles that approximate statistical mechanics to predict the future based on that ensemble information. However, there have been relatively few studies of how we conceptualize the motion and physics of a large set of objects (aside from judging direction of motion in random dot kinematograms). It is therefore an empirical question of whether we can (a) extract ensemble physical properties like energy, momentum, or acceleration, and (b) use that information to extrapolate the future state of the world. Understanding how we might compress our representations of sets of objects to efficiently predict the motion of multiple objects is therefore key to understanding how we represent physical rules in real-world scenarios.

Choices of simulation contents

When initializing a simulation, an important consideration is what parts of the world to include in the simulation. While most experimental studies make this a trivial problem – the relevant information is given on the computer screen or on the page in all of the experiments in this thesis – it is not clear how we set this criteria in our day-to-day lives when the relevant information is less obvious. If you are throwing a ball to a friend in the middle of a field, for instance, it seems unnecessary to consider how a tree fifty feet to the side might affect the trajectory of the ball. On the other hand, if that tree is just a few feet to the side, it would be important to consider that the ball might hit an overhanging branch. How then might we choose when it is appropriate to include an object in our simulations, and when it is unnecessary?

One possible consideration is proximity: if you are throwing a ball to your friend, then we can a priori determine that the scale of simulation is somewhat larger than the distance between you and your friend, and therefore all objects within that distance might be considered important for simulation. On the other hand, if you are bowling, the relevant distance would be that of a bowling lane, or if you are trying to pour coffee, the relevant scale would be the distance between the carafe and the coffee cup. Under this account, an area is selected based on the ultimate goal of simulation, and all objects within that area are included in the simulation.

Another possibility, however, is that objects are dynamically added to the simulations. Perhaps we start by simulating the bare minimum required to achieve our goals: just your friend and the ball if you are playing catch. This simulation can then be checked against objects in the world, and if the path of the ball would intersect with or pass nearby other objects (e.g., a tree), that object is added to future simulations. While this process incurs more work to check simulations against the world, it can also save on the effort of simulating the motion of objects that will in the end not affect our greater goal.

However, while these proposed approximations will carve out a section of the world for simulation, they both would miss far-away moving objects that will later intersect with the path of the important objects. If you are bowling, the apparent scope of simulation might be your bowling lane, but if someone to the side throws the ball wildly and it is bouncing across a number of lanes, it is important to consider how it might impact your toss. This leads to a circular question: how do you know that the ball will cross your lane without simulation, but how do you know that it is important to simulate it until you theorize it might cross your lane? It is therefore an open question of how we might determine outside moving objects, or whether we even can in the absence of something to draw our attention to them.

The extent of extrapolation

Our simulations cannot go on forever; instead there must be a limit to how far we can extrapolate the motion of objects (either in time, or in distance). This problem itself encompasses two core unknown features about simulations: how far into the future do individual simulation paths go, and how far into the future can we push our predictions.

In all cases of physical simulation, we must determine when to end any given simulation path. With certain goal directed simulations, there is a natural endpoint – for instance, when the ball crosses the plane of the bucket or paddle in Chapters 2 & 3, or when it reaches one of the two targets in Chapter 4. But even this cannot be a hard rule, since an individual simulation path might wander for a long time before reaching a goal. And this becomes more difficult for less constrained problems where an obvious endpoint does not exist. If we want to know where a moving object will come to rest, for instance, it is impossible to a priori know how far or how long it will travel. To avoid wasteful simulation, we might therefore cut off simulations that are less likely to provide us with relevant information. Perhaps we set a time cutoff, so that simulations that take too long to provide us with useful information are thrown away, but this leaves open the question of how we decide what this time cutoff should be. Perhaps this cutoff is tied to the choice of simulation contents, so that if an important object leaves the relevant area, we cannot extract useful information from that simulation and therefore ignore it. Or perhaps this is just an inherent limit of our cognitive machinery – there is only so far into the future that we can drive a simulation, but this will differ depending on the complexity of the situation. Determining how our individual simulations are limited is therefore an important component to understanding the process of physical reasoning.

At a higher level, we can only predict the motion of objects so far into the future. This is of course limited by how far our simulations can run, but our predictions might fall apart before any limits of individual simulations. Another limitation in our predictions might be the variability between simulations – if uncertainty about where any object will go is too great beyond a certain distance, we would do just as well to say that we have no idea of its future position. For instance, if you were to be asked about where an object falling down a long set of stairs would end up, it is easy to predict the motion of a solid box (it will likely just tumble to the foot of the stairs), but trying to predict the same for a rubber ball would be nearly impossible, since there is so much noise in the way that it might bounce. Yet if you were asked to predict that rubber ball's motion after two steps, this might be difficult but doable. Studying the amount of uncertainty that can accumulate before prediction becomes impossible will help set limits on how far into the future we can set our predictions using simulation.

Temporal resolution of simulations

A final important consideration in physics simulations is how frequently the world is updated: if there is only a very short period over which the world is updated between checks for interesting events (e.g., collisions) then many needless checks will be made, but with too much time between checking for events where we approximate object motion with ongoing dynamics we might miss something important (e.g., two objects moving perpendicular then colliding).

Determining how much time should pass between resolving collisions is an important part of a computer physics engine. Many physics engines use a fixed timeframe for making updates for computational simplicity (Millington, 2010). However, this is not the only way to set the temporal resolution of a simulation engine – for instance, two-phase collision detection algorithms check quickly if objects' trajectories will bring them close to each other, then slowly check for actual collisions (Mirtich, 1997), while time-division engines will try to find the moment of first contact between two objects rather than working at a fixed scale (Millington, 2010). The many options available for computer physics engines suggests that there are many possible ways that the mind might treat the temporal resolution of its physical simulations.

The method of determining how far apart these detections are spaced will affect the amount of physical approximation required between these steps, and therefore the potential errors that might arise from simulation. Therefore, studying the temporal resolution of our approximate physics engines – whether that resolution is fixed or variable, and how it is determined – is important for understanding both the amount of effort underlying simulation and the types of approximation errors we might make.

5.2.2 Approximations to probabilistic reasoning

I have also argued that physical reasoning accounts for various sources of uncertainty, and this uncertainty causes our predictions about the future to be probabilistic, weighting various outcomes by how likely they are to occur. For instance, Smith and Vul (2015) demonstrate that our confidence in our physical predictions is correlated with the variability in predictions across people, and so it is likely that this well-calibrated meta-knowledge of our own uncertainty is formed from a probabilistic distribution over possible future world states.

However, producing a complete, continuous probability distribution over all possible future world states is computationally intractable. Because our physical updating is dynamic (Chapter 3), the only way to produce a single possible future is to iteratively update our simulations. Thus most recent physical simulation work (including this thesis) has used *Monte Carlo simulation* to extrapolate the future under uncertainty – this forms a posterior probability distribution by sampling potential current states of the world (weighted based on the present uncertainty), then running the simulation based on those starting conditions and tallying how often the simulations end up in a given state (e.g., where the ball would cross the plane of the bucket in Chapter 2, or whether the ball landed in the 'red' or 'green' goals in Chapter 4). In the limit of taking an infinite number of samples, this would produce the exact posterior probability of a physical event occurring, but even with a limited number of simulations, this probability can be approximated.

Of course, it is impossible to create an infinite number of different simulations within a finite time. Instead, the cost of taking each additional sample (in mental effort required to produce that sample or opportunity cost of not acting) must be weighed against the expected benefit of the additional information that sample might provide (Vul, Goodman, Griffiths, & Tenenbaum, 2014). If producing a simulation is cognitively effortful and the benefit of making the correct prediction is low (for instance, in the middle of a psychology experiment where there is no remuneration for accuracy), then people may take only a small number of samples before deciding to act on that information. Indeed, Hamrick et al. (2015) suggest that when presented with a dichotomous prediction (whether a ball will travel through a hole in a wall or not), people perform simulations only until there is a net evidence of just two samples in favor of one of the options. Thus the probabilistic machinery underlying physical reasoning may be based on extremely coarse approximations.

But if people use very few simulations for physical reasoning, then how can this explain the constant updating of predictions observed in Chapter 4? Although peoples' behavior was consistent with producing new simulations on a regular basis, most of the time an additional simulation would provide little additional information: the probability of a ball reaching one target is unlikely to change much if all the ball has done is moved slightly along a straight line. Instead, simulation outcomes generally changed only after what we would consider to be an event – for instance, the ball bouncing off of a wall, or narrowly missing a wall we expect it to hit. Therefore, to save effort, the mind might produce new simulations only when it expects the outcome to be different from the past simulations – after a collision, or when prior simulations are no longer good descriptions of the current state of the world, for instance.

A hint of how this might be possible comes from approximations to probabilistic reasoning within the domain of machine learning. Many object tracking programs in computer vision must account for the fact that their localization of objects within each picture frame will necessarily be noisy or unknown due to occlusion. Therefore, to gain certainty in object localization, these programs use not just current noisy observations, but also the expected position of the object from prior observations and associated dynamics. Because this is a computationally difficult problem, it is often approximated using *particle filters* – proposing a limited number of possible object positions (particles), extrapolating their motion, then comparing those expected positions to noisy observations and allocating more belief to those particles that are closer to the noisy observation (Doucet, De Freitas, & Gordon, 2001). This algorithm has had success both in computer tracking algorithms (Okuma, Taleghani, De Freitas, Little, & Lowe, 2004), and has been proposed as a model for how people track multiple objects (Vul, Frank, Alvarez, & Tenenbaum, 2009).

The motion extrapolation required for this tracking is very similar to the noisy Newton framework, albeit at a more limited scale. If people do track objects by keeping in mind a limited number of plausible locations and where those objects will go, then predicting motion over time might rely on a similar process: proposing a handful of plausible trajectories for how an object will move over time, and allocating belief to those trajectories based on how well they describe the observed motion as the world unfolds. This might explain why our predictions change after a ‘surprising’ event without requiring constant re-simulation: typically, our predictions will be good enough that we have no need to form new proposed trajectories, but if an event produces object motion that is very different from the prior proposals, we will decide that our past predictions are no longer good enough and need to form new simulations to update our predictions. While this is just one proposal for how we might efficiently approximate probabilistic physical reasoning, we can study how we save effort in physical simulation and propose other approximations by tying together theories from machine learning and psychology.

Physical reasoning is a domain in which updating our predictions in concert with changes to the world is a core requirement. But physical simulation is computationally taxing, making approximations all the more important. Studying how people accomplish this complex task both efficiently and effectively can provide insight into the general algorithms that the mind uses to approximate probabilistic reasoning.

5.3 Conclusion

Physical reasoning pervades our daily experiences so much that we barely notice it, yet relies on complex mental operations to accomplish. This thesis provides an initial framework that explains how people use this reasoning to understand and interact with the world: when we engage with the world we support our reasoning with relatively accurate physical simulations, given uncertainty about the properties of the objects we simulate. Fully describing human physical simulation will require significant research into the underlying cognitive processes and neural bases, guided by knowledge of computer physics engines as plausible algorithms for the underlying simulations. To expand this knowledge to the whole of physical reasoning will require studying how physical simulation and explicit knowledge interact with one another, and how we decide which system of reasoning to use. But despite the long road ahead, understanding this process will allow us to explain a core facet of the human experience and provide insight into the processes the mind uses to build an understanding of both the current environment and the future.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science*.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487.

Catrambone, R., Jones, C. M., Jonides, J., & Seifert, C. (1995). Reasoning about curvilinear motion: Using principles of analogy. *Memory and Cognition*, *23*(3), 368–373.

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 30.

Davis, E., Marcus, G. F., & Chen, A. (2013). Reasoning from radically incomplete information: The case of containers. *Advances in Cognitive Systems*, *2*, 273–288.

Doucet, A., De Freitas, N., & Gordon, N. (2001). *An introduction to sequential monte carlo methods*. Springer.

Fischer, J., & Kanwisher, N. (2015). The neural basis of intuitive physical reasoning. *Journal of vision*, *15*(12), 518–518.

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Towards a theoretical framework. In R. Michalski, J. Carbonaell, & T. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, p. 311).

Frick, A., Huber, S., Reips, U.-D., & Krist, H. (2005). Task-specific knowledge of the law of pendulum motion in children and adults. *Swiss Journal of Psychology*, *64*(2), 103–114.

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.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753.

Hamrick, J., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? Optimal mental simulation tracks problem difficulty. In D. C. Noelle et al. (Eds.), *Proceedings of* *the 37th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Helm, H., Gilbert, J., & Watts, D. M. (1985). Thought experiments and physics education. *Physics Education*, *20*(5), 211.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 795–803.

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

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, *107*(25), 11163–11170.

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

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

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

Millington, I. (2010). *Game physics engine development: How to build a robust commercial-grade physics engine for your game*. Boca Raton, FL: Taylor and Francis.

Mirtich, B. (1997). Efficient algorithms for two-phase collision detection. *Practical motion planning in robotics: current approaches and future directions*, 203–223.

Nealen, A., Müller, M., Keiser, R., Boxerman, E., & Carlson, M. (2006). Physically based deformable models in computer graphics. In *Computer Graphics Forum* (Vol. 25, pp. 809–836).

Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., & Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In *Computer vision-eccv 2004* (pp. 28–39). Springer.

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

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

Shanon, B. (1976). Aristotelianism, Newtonianism and the physics of the layman. *Perception*, *5*(2), 241–243.

Smith, K. A., & Vul, E. (2015). Prospective uncertainty: The range of possible futures in physical predictions. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

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

Vul, E., Frank, M. C., Alvarez, G. A., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22).

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

White, P. A. (2012). The impetus theory in judgments about object motion: A new perspective. *Psychonomic Bulletin and Review*, *19*, 1007–1028.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences*, *106*(15), 6088–6093.

A Supplemental materials for Chapter 2

A.1 Model details

Here we discuss the mathematical formulations behind the model predictions of physical reasoning. The models were split into two components. The first component is a *forward model* that can be thought of as a function $R(t_{rel}, y)$ which returns the predicted position where the ball would cross a line at height y if released at time t_{rel}. Note that this function does not in of itself make any claims about how prediction works, but can instead be set for any assumption about how people extrapolate the motion of the ball. The next component, the *task action*, uses this function to determine where to place the bucket in the catching task, or when to release the ball in the releasing task.

A.1.1 Forward models

Noisy Newton

The noisy Newton forward model captures the position and velocity of the ball at the moment of release, and uses Newtonian ballistic motion equations to extrapolate the path of the ball:

$$
\begin{aligned}
x(t) &= x_0 + v_{x0} * t \\
y(t) &= y_0 + v_{y0} * t + \frac{g * t^2}{2}
\end{aligned}
\tag{A.1}
$$

The forward model then returns the *x* position of the ball when it reaches the vertical position of the bucket.

However, people must also judge the depth of the pendulum behind the computer monitor. The inverse relationship between cord length and gravity given a constant pendulum period means that estimating depth (and thus pendulum length) with constant gravity is mathematically equivalent to assuming a constant cord length and estimating the force of gravity. Because responses were measured using a constant unit of on-screen distance (in pixels), for computational efficiency the noisy Newton model assumed a constant cord length and estimated the effective strength of gravity in $\frac{px}{s^2}$.

Non-physical models

Each of the non-physical models assumed that the ball would travel in a straight line from its release point as an angle away from the vertical (θ_r). Each of these release angles was calculated as a function of the angle the pendulum cord made with the vertical at the moment of release (θ_c).

The *angled* forward model calculated the release angle as a piecewise linear function of the cord angle. There were two intercepts and two slopes for this function, so that the ball would travel differently depending on whether it was swinging downwards or upwards:[1]

[1] These angles were mirrored when the ball was traveling leftward for symmetry

$$\theta_r = \begin{cases} i_1 + s_1 * \theta_c & \text{if } \theta_c > 0 \\ \\ i_2 + s_2 * \theta_c & \text{otherwise} \end{cases} \tag{A.2}$$

The *outward* model assumed that the ball would continue along the path of the cord, but allowed for the angle to shift upon release. Thus the ball angle was calculated to be the same as the release angle, with an adjustment:

$$\theta_r = a * \theta_c \tag{A.3}$$

The *straight down* model simply assumed that the ball would drop upon release:

$$\theta_r = 0 \tag{A.4}$$

Once the path of travel was calculated, the model predicted the landing position of the ball as the horizontal position of where the path line intersected the plane of the bucket.

A.1.2 Task actions

Catching

We assume that the best location to place the bucket to catch the ball would be where the forward model suggests the ball will land. However, people must account for both motor and extrapolation error, which we formalized as Gaussian noise. This noise increased linearly with the vertical distance between the bucket and release height of the ball (h_{tr}), and we fit two free parameters to capture the slope and intercept of this relationship:

$$\sigma_{tr} = a_c + b_c * h_{tr} \tag{A.5}$$

Thus the choice of where to place the bucket on the catching task (S) on a specific trial can be described by:

$$S_{tr} \sim \mathcal{N}(R(t_{tr}, y_{tr}), \sigma_{tr}) \tag{A.6}$$

Releasing

In the releasing task, we assumed that people would have a reasonable sense of where the ball will go if released at each point in time. Similarly, the model can use its forward predictions to determine at each point in time where the ball would land if released from the cord.[2] From this information, we can form a function over possible release times that returns 1 if the model will land in the bucket, and 0 otherwise:

$$L(t)_{tr} = \begin{cases} 1 & \text{if } R(t, y_{tr}) \in \text{bucket}_{tr} \\ 0 & \text{otherwise} \end{cases} \tag{A.7}$$

The optimal time to release the ball (T_{dec}) was assumed to be the middle of any contiguous time period in which $L(t) = 1$. If there were two optimal release times (T_1 and T_2), the model probabilistically chose one, preferring the point with the smallest vertical distance to the bucket. This choice was formalized as a logistic function on the difference between the ball heights at each point (h) with a single scaling parameter (s_r), but no intercept shift (since we assumed that at equal heights, participants should be ambivalent about which time to choose):

[2] This was approximated analytically by determining $R(t, y_{tr})$ for all t segmented in blocks of 10ms.

$$p(T_1) = \text{logistic}(s_r * (h_{T_1} - h_{T_2})) \tag{A.8}$$

Once the model chooses the time point that it aims to cut, its actual release time for a trial (T_{rel}) was selected as value from around that choice with Gaussian noise fit as a free parameter (σ_{terr}) to reflect the motor errors that people make:

$$T_{rel} \sim \mathcal{N}(T_{dec}, \sigma_{terr}) \tag{A.9}$$

A.1.3 Parameter fitting

For all models, parameter fitting was accomplished by maximum likelihood estimation on individual trial data. For a given set of parameters, the models determine the mean and variance of bucket positions for the catching condition, and of release timing in the releasing condition. Because forward model parameters were shared across both tasks, we calculated the likelihood of a given set of parameters as the combined likelihood of all responses for both the catching and releasing tasks. Responses were aggregated over participants for general model comparison, but split by participant for the individual model fitting.

A.2 Supplementary results

A.2.1 Exclusion principles

To ensure that we were measuring the physical principles that people use when interacting with the world, we want to exclude participants who were not actually carrying out the task (due to inattention or laziness). To determine whether participants were making an effort, we asked how reliably their responses on the catching and releasing

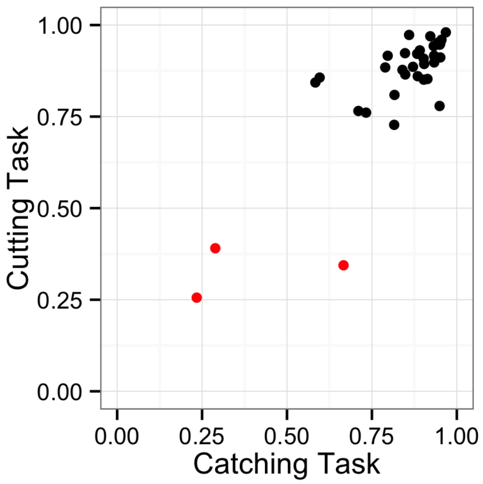


Figure A.1: Measures of the coefficient of determination for each participant in both the catching and releasing tasks. Each point represents a single participant. Three outliers (marked in red) were eliminated because those participants demonstrated poor response differentiation by trial.

tasks were affected by trial type. We measured this as the coefficient of determination (R^2) of an ANOVA predicting each individuals catching and releasing responses as a function of the 48 trial types, which measures how much differentiation by trial exists. Most participants responses varied significantly by trial, with high average coefficients of determination (catching: $R^2 = 0.83$; releasing: $R^2 = 0.84$); however, there were three participants whose responses were significantly less differentiated than all others (catching $R^2 = [0.23, 0.29, 0.67]$; cutting $R^2 = [0.26, 0.34, 0.39]$; see Figure A.1) we eliminated those three participants from all other analyses. Because this metric did not make any assumptions of how participants varied by trial, but rather only if they varied, we viewed this as an appropriate exclusion criterion that would not a priori favor any of the possible models, accounts, or hypotheses.

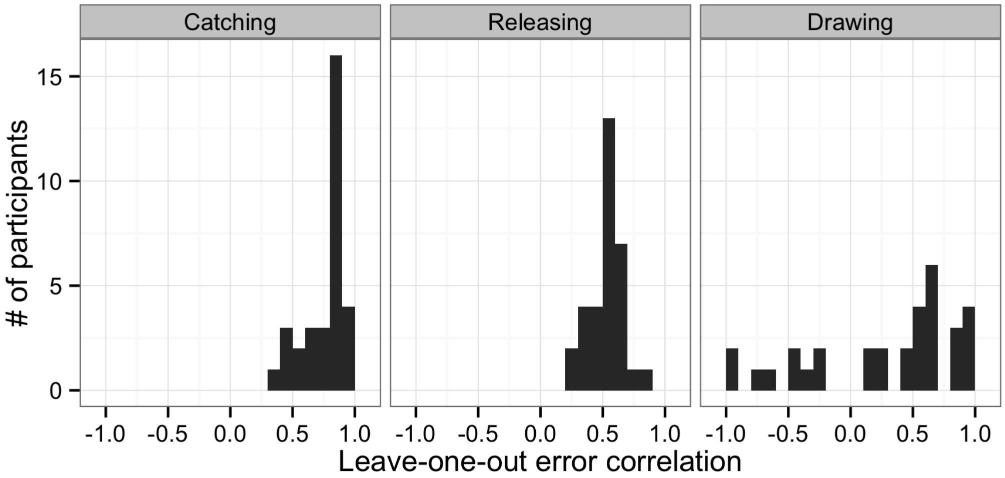


Figure A.2: Histogram of error correlations between each subject and all other subjects across all three tasks in Experiment 1. There was low variability in error correlations on the catching and releasing task, while the drawing error correlations suggested idiosyncratic reasoning.

A.2.2 By-subject, by-task error correlation

In order to measure the consistency of responses across participants, we reported the correlation between individual participants' errors and the errors of all other participants. This was trivial to calculate for the catching and releasing tasks; however, for the drawing task we generated matched data by extrapolating drawn trajectories to produce equivalent 'bucket positions' as if they were guided by their drawings.

As can be seen in Figure A.2, any participant's responses on the catching or releasing tasks could be well predicted by how all other participants responded in Experiment 1. However, participants were significantly more variable on the drawing task, suggesting that these conceptual responses were in general more idiosyncratic.

Drawing extrapolations from Experiment 2 show similar variability to the drawings from Experiment 1, as can be seen in Figure A.3. With all stimuli included, participants from the Static condition are more correlated with others from the *Static*

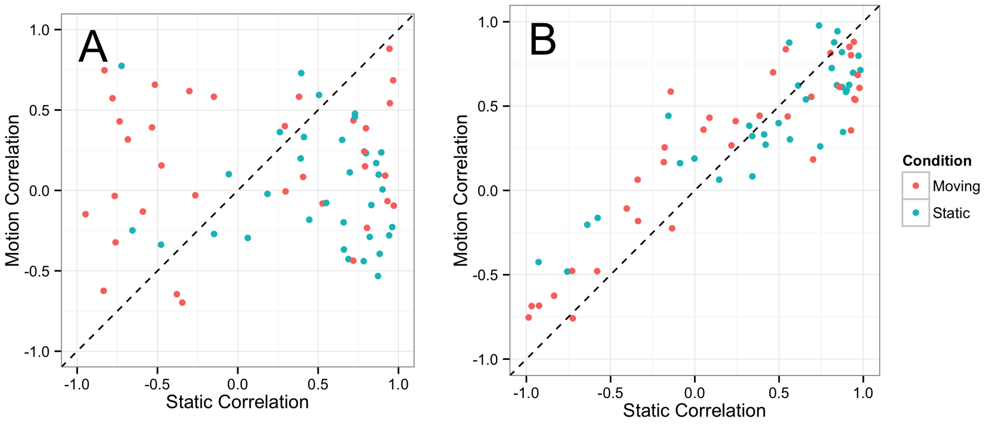


Figure A.3: Plots of correlations between each participant's extrapolated drawings and the average of all other subjects (split by condition) from Experiment 2. *A:* Including all stimuli. *B:* Excluding the nadir stimulus.

condition, and vice versa for the Motion condition (in Figure A.3A, *Static* participants tend to be below and to the right of the identity line, which *Motion* participants tend to be above and to the left). However, this higher correlation for one's own condition is driven almost entirely by the nadir stimulus.

A.2.3 Learning

Although we hid all ballistic trajectories with the exception of one sample trial each in the catching and releasing tasks, it is possible that participants were learning from the binary feedback provided, perhaps by trial and error. Learning the correct response pattern could bias the results to be more 'Newtonian' even if knowledge is based on simple model-free predictions.

We measure learning as decreasing error (relative to ground truth) over the course of the experimental session. There is no evidence that the average participant was learning in either the catching ($b = -0.0028$ cm/trial, $t(16.2) = -1.39, p = 0.18$) or releasing ($b = -0.0010$ cm/trial, $t(30.2) = -1.03, p = 0.31$) tasks; over the course of the 240 trial experiment, this would amount to just a 0.66cm decrease in error on the catching task (15.7% of the average error), and 0.25cm decrease in error on the cutting task (7.3% of the average error). Therefore it is unlikely that people were behaving in line with the noisy Newton model due simply to model-free learning.

A.2.4 Individual model fitting

Twenty-eight of thirty-two individual participants were best fit by the Newtonian model rather than an alternate. Here we report how good these individual fits were. As can be seen in Figure A.4, the participants whose predictions were better explained by alternate models were only marginally better explained, whereas for most subjects the Newtonian model explained behavior significantly better than any alternatives.

A.2.5 Predicted response variability

The task action assumptions directly influence the amount of response variability expected by each model. Therefore a good test of whether these assumptions are reasonable is to determine whether they predict the trial-by-trial variability in an unbiased model. Because the noisy Newton model is nearly unbiased, we used it as a comparison point. The trial-by-trial variability predicted by the model was well correlated with participants actual variability (catching: $r = 0.79$, releasing: $r = 0.82$), suggesting that the task actions posited for the model do a reasonable job of capturing how participants were performing each of the two tasks.

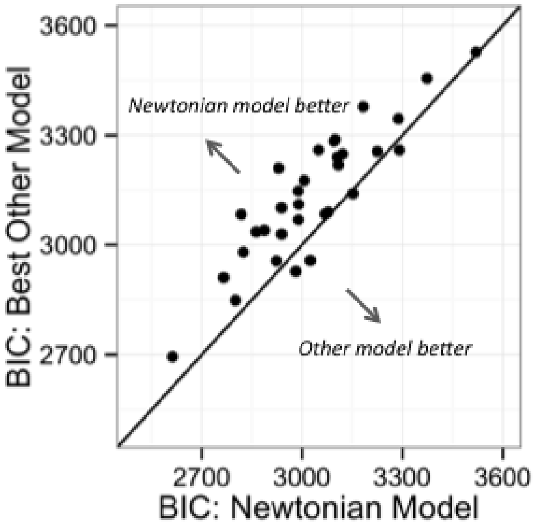


Figure A.4: Individual model comparisons. Each point represents a separate participant, comparing the BIC of the noisy Newton model fit to their data (x-axis) versus the BIC of the best alternate model (y-axis). Because lower BIC suggests a more parsimonious model, participants above and to the left of the identity line have responses characterized best by the noisy Newton model, while those below and to the right are best characterized by an alternate model. In general, the noisy Newton model characterized responses much better than alternatives.