On the Nature and Origin of Intuitive Theories: Learning, Physics and Psychology

by

Tomer David Ullman

B.Sc., Physics and Cognitive Science, Hebrew University (2008)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Signature redacted

Author ....
Department of Brain and Cognitive Sciences
January 15, 2015

Signature redacted

Certified by ....
Joshua B. Tenenbaum
Professor
Thesis Supervisor

Signature redacted

Accepted by ....
Matthew Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar,
Director of Graduate Education for Brain and Cognitive Sciences

On the Nature and Origin of Intuitive Theories: Learning, Physics and Psychology

by

Tomer David Ullman

Submitted to the Department of Brain and Cognitive Sciences
on January 15, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

This thesis develops formal computational models of intuitive theories, in particular intuitive physics and intuitive psychology, which form the basis of commonsense reasoning. The overarching formal framework is that of hierarchical Bayesian models, which see the mind as having domain-specific hypotheses about how the world works. The work first extends models of intuitive psychology to include higher-level social utilities, arguing against a pure 'classifier' view. Second, the work extends models of intuitive physics by introducing a ontological hierarchy of physics concepts, and examining how well people can reason about novel dynamic displays. I then examine the question of learning intuitive theories in general, arguing that an algorithmic approach based on stochastic search can address several puzzles of learning, including the 'chicken and egg' problem of concept learning. Finally, I argue the need for a joint theory-space for reasoning about intuitive physics and intuitive psychology, and provide such a simplified space in the form of a generative model for a novel domain called Lineland. Taken together, these results forge links between formal modeling, intuitive theories, and cognitive development.

Thesis Supervisor: Joshua B. Tenenbaum
Title: Paul E. Newton Career Development Professor

Acknowledgments

A group of lions is a pride, a pack of rhinos is a crash. What is the right term of venery for a bunch of acknowledgments?

I have an *embarrassment* of acknowledgments:

My singular advisor Josh Tenenbaum, for his probing insights, his wonderful intellect, and his incredible dedication.

Laura Schulz, for her encouragement and inspiration, and for being a sporting adversary in the ongoing Ullman vs. Schulz affair.

Noah Goodman, for being an unofficial fountain of sage advice and then an official one, and for lodging his voice in my head as that of reasoned skepticism.

Liz Spelke, for helping me realize what I want to be when I grow up.

Whitman Richards, for his extraordinary thoughts and comments.

Frank Jäkel, for his friendship near and far.

John McCoy, for his peerless company, and for his kindred interest.

Brenden Lake, for not hating me after the 'incident', and the good times that followed. We will always have Nara.

Andreas Stulmüller, for his axiomatic and well-defined friendship.

Owen Macindoe and Sophie Mackey, for some serious levity.

Tobi Gerstenberg, for his humor and charm, and for reminding me of that scene from the Big Lebowski.

Liz Bonawitz, for antics, shenanigans, and sing-alongs.

Steve Piantadosi, for his amiable tolerance, and for providing a role model of good computational psychologist.

Celeste Kidd, for her cheerful enlightening conversations, and for maggot cheese.

Jon Malmaud, for cryptic witty exchanges.

I have a truly marvelous list of people to thank, which the margin is too small to contain

My friends and colleagues in CoCoSci-land, broadly defined, for their wit and wisdom, their comments and camaraderie: Chris Baker, Eyal Dechter, David Wingate, Peter Battaglia, Virginia Savova, Jessica Hamrick, Tim O'Donnell, Chris Bates, Tao Gao, Max Siegel, Max Kleiman-Weiner, Mark Velednitsky, Sam Zimmerman, Joshua Hartshorne, Leon Bergen, Rus Salakhutdinov, Dan Roy, Yarden Katz, Cameron Freer, Anna Rafferty, Joseph Jay Williams, Chris Lucas, Michael Pacer and Sam Gershman.

My friends and colleagues in the developmental world, for being welcoming guides in terra incognita: Hyo Gweon, Kim Scott, Yang Wu, Paul Muentener, Pedro Tsividis, Rachel Magid, Melissa Kline, Julia Leonard, Julian Jara-Ettinger, Samantha Floyd, Daphna Buchsbaum, Kiley Hamlin and Alison Gopnik.

The multitudes who participated in my experiments over the years, for the wisdom of individuals.

My family, for everything.

Contents

List of Figures

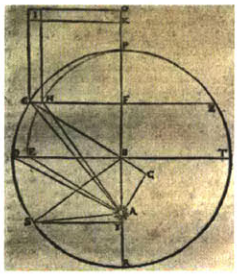
Chapter 1

Introduction

"The only innocent feature in babies is the weakness of their frames; the minds of infants are far from innocent." — Augustine of Hippo, Confessions

---

Even before we know the world, we know *about* the world. From birth, we have expectations about objects, magnitude, space and action. This core knowledge forms our basic intuitions. And yet, cognitive science does not have a formal theory of these basic intuitions. These are not trivial statements, they are hard-won recognitions established over the past decades through experimental work with infants, children and adults. By looking at what infants find surprising and what they prefer, researchers amassed a wealth of knowledge about what infants expect and know: objects follow smooth paths and don't wink in and out of existence; agents act efficiently to achieve goals; numbers can be added and subtracted, and so on. Despite these general principles, there is no explanatory computational theory to unite and explain



"Ellipsin fieri orbitam planetæ" (Kepler, Epitomes astronomiae Copernicanae)

the separate strands of findings. The state of the field resembles pre-Newtonian astronomy, a period when people rigorously collected a copious amount of data and expressed general qualitative principles about heavenly motions, but lacked a formal quantitative and principled account. The difference between data-based generalizations and formal theory is the difference between saying "Planets follow elliptical paths with the sun at a foci" and "$F = m \cdot a$ and gravitational works in an inverse square way, therefore the planets will move *thus*".[1]

At about the same time that the 'core knowledge' account of infant knowledge was crystallizing, computational cognitive science was developing new ways to think about thinking. Structured generative models emerged as influential tools for capturing the computations of the human mind. Following the theory-based approach in cognitive science [123], these new tools view the mind as reverse-engineering the world, searching for theories and causes that explain perception. In the pre-Newtonian era we find ourselves, this formalism is a bit like calculus: an important computational advance in itself, but hard work is needed to link it up with the real world.

In this dissertation, I present several such links between computational theories and intuitive theories. The dissertation is concerned with the common questions of researchers in both AI and development, namely representation and learning, or "what we know" and "how we get more of it". On the question of representation, I focus on the core domains of intuitive physics and intuitive psychology, and the connections between them. On the question of learning, I propose that many learning challenges are best addressed at the algorithmic level of modeling, and suggest such an algorithm, drawing parallels between the dynamics of the algorithm and the

[1]This is not for lack for trying. There have been attempts to formalize cognitive development, as the historical section shows, and these attempts are ongoing.

way children learn. Throughout the dissertation I present empirical, theoretical and philosophical support for the particular claims put forward. But I also allow myself to speculate on what models *should* be like, with the hope that the reader will forgive or even enjoy such speculations.

The rest of the introduction is meant to equip the reader with the background, terms and details necessary for their journey through the thesis itself. My hope is that by the end of the introduction, the reader will be able to answer for themselves on a basic level: "What is the relationship between cognitive modeling, intuitive theories and cognitive development? What do we know about child development and intuitive theories today, and what is a good formal account of that? What's the alternative?"

I first review the historical exchange between computational models and cognitive development (Section 1.1). Next, I describe current influential views in development including "Core Knowledge" and the "Theory Theory" (Section 1.2), and broadly what we think infants know about the core domains of agents and objects. Building on this, I ask what are the criteria for a formal account of infant core knowledge in principle. In Section 1.3 I give an overview of a formalism that matches these criteria: hierarchical Bayesian models (HBMs), and explain their connection to intuitive psychology and intuitive physics. Section 1.4 introduces the distinction between a computational level and an algorithmic level analysis, and uses the distinction to explore an oft-cited criticism of HBMs: Even assuming these models get the representation right, how can they learn anything truly 'new'? Finally, Section 1.5 presents an approach based on cues, features or rules, that will serve – in various guises – as the main foil for the HBM account.

Roadmap of intro

Also, here is a brief outline of the structure and contributions of the next chapters: Chapters 2-3 focus on the core domains of agents and objects. Chapter 2,

Roadmap of thesis

motivated by experiments with pre-verbal infants, extends a formalism that models action understanding as 'inverse planning' to include social goals such as helping and hindering, provides strong evidence against a cue-based account and argues for an innate or early-developing mentalistic apparatus. Chapter 3 builds on the proposal that intuitive physics is based on a mental 'physics-engine', asking: what parts of this engine can be learned, and how? Chapter 4 tackles the question of learning, and proposes that by focusing on the algorithmic level of structured generative models – particularly on stochastic search algorithms – we can address several philosophical and psychological puzzles about how children learn. Finally, Chapter 5 examines the challenge of cross-core-domain connections, going back to agents and objects and proposing a generative account of how people reason when common-sense explanations require understanding something about both psychology and physics.

1.1 Formal Models and Child Development, a Brief History

Formal modeling of what children know and how they develop is not a new suggestion. The changes in the field of artificial intelligence have often paralleled, influenced, and were influenced by changes in the field of child development. This is hardly surprising, as both fields are mainly concerned with the representation and acquisition of knowledge: What it is, and how we get more of it.

Even before the proposed equivalence of the mind with computation – the 'driving metaphor' of the field of cognitive science [134] – researchers in the nascent fields of child development and computation were seeing parallels. Prior to the 'cognitive revolution', Piaget was examining the child's mind in terms of logical symbols, mental

models and mental operations [129]. Around the same time, Turing suggested that rather than simulating the adult mind, we might be better off trying to recreate the mind of the child and teaching it so as to produce the adult mind [182].

We cannot expect to find a good child machine at the first attempt
– Alan Turing

During the mid 20th century, researchers in proto-AI and psychology were struggling with similar questions: How much knowledge is there at the beginning? How is knowledge represented? How is new knowledge learned?

On the question of the initial state of knowledge, Turing considered the child's mind a notebook with 'rather little mechanism, and lots of blank sheets' [2]. In this, Turing's outlook was in many ways similar to the dominant behaviorist view in the United States at the time, positing little initial structure and thinking that some rewarding or punishing signal would allow the child program to correctly learn new knowledge [182]. Turing, Piaget and Skinner could all be seen as similar in their relatively empiricist belief (or hope) that the initial state of the child is close to a blank notebook / sensorimotor machine / unconditioned subject. Such a view contrasted with the ideas researchers like Chomsky [30], who argued for the innate existence of conceptual content (such as grammatical rules).

Give me a child, and I'll shape him into anything
– B.F. Skinner

As for the question of how knowledge is represented, Turing (and constructivists like Piaget) suggested the child program discovers some formal structure, a subprogram or set of mental operations. Such mental, inner structures were denied by the behaviorist tradition.

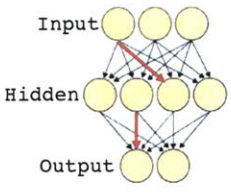
Finally, regarding the question of new knowledge, computational models were called upon early on to address this challenge, be it models of operant conditioning, assimilation or schema transformation [162, 129]. It is interesting – though not

[2]Or rather, Turing 'hoped' this was the case, as it would be much easier to program such a machine, and perhaps anticipating the difficult task of uncovering innate structures should those exist.

surprising – that a "short blanket" problem occurred when trying to solve both the issues of representation and learning. A short blanket covers either the head or the feet, but not both. Simple learning rules, such as the Rescorla-Wagner learning rule, were easy to implement and study, but could not account for rich knowledge [136, 120]. Rich knowledge, captured by representations such as grammar, was either assumed as given, or was not provided with an implementable formal treatment (e.g. Piaget's theories [129]).

During the rise of the cognitive sciences, the information processing approach to modeling was highly influential on cognitive development [98, 161]. When the field of cognitive science was focusing on symbolic logic [126], learning was seen as the acquisition of new 'rules' for reasoning about domains. Much as an intelligent program could acquire new 'rules' for achieving a goal-state in a toy-world, children and adults were modeled as learning new logical-rules, and experiments on explicit rule learning became popular [161]. Both children and adults were modeled as acquiring new rules within a production system, but development was seen as a program-transformation going from one production system to the next [161]. This view suggested two types of programs necessary to describe development: many 'stage-programs' that captured the mental state of a child at each developmental stage, and one 'transformation program' that takes as input a stage-program and outputs a different stage-program. This distinction, made by Simon, was influenced both by Turing (the mind as a program) and by Piaget (separate stages of development).

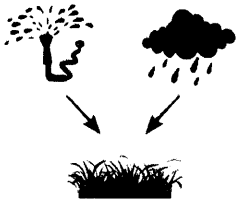
Perhaps we would settle for a theory of something less than the whole child
– Herbert Simon



With the advent and popularization of connectionist architectures in the 1980's, it was proposed that development was the simple ongoing process of adaptive weight change in a neural network. What was previously seen as the discrete acquisition of a new basic understanding of some concept or the relationship between concepts [114, 139, 117, 138]. (e.g. "If there is more weight on one side, an apparatus will

tilt towards that side"), was now seen as coming about through quick and drastic weight shifting when enough data was supplied. Nowhere in the network was there an explicit concept, or rule, or transformation. Development and learning were now, to some degree, equivalent. Much as the network can adjust weights to learn a new word in French, it can adjust weights to recognize new objects, or to 'realize' that both weight and distance are important in predicting the movements of a balance scale.

Around the same time that parallel distributed processing was coming into fashion, both developmental researchers and AI researchers became concerned with questions of causal reasoning and uncovering the 'true' structure of the world. In AI, Judea Pearl was developing Causal Bayes Nets [127], which aim at modeling the underlying structure that led to an observation, rather than just finding correlations between observations. Causal Bayes Nets do this by combining explicit predicates (such as 'symptom' or 'disease') with probabilistic Bayesian inference, and with a notion of 'intervention' that isolates causal influences. Independently of this research, researchers in development were proposing that children concern themselves with finding the 'true' underlying structure of events, building theories and revising them like scientists [69]. This 'theory theory' view is discussed in the next section, and the historical review is far too brief to do justice to the Causal Bayes Net approach (much as it is short on justice towards connectionism). For my purpose, the takeaway is that Pearl's research on causality had an important influence on developmental research beginning in the late 90's [65, 68], when some researchers in computation and development began seeing Causal Bayes Nets as the formal nuts and bolts of the theory-theory. Rather than learning a "rule" relating predicates, or adjusting weights in a connectionist network, children were seen as distinguishing, comparing and choosing among different causal networks for explaining a situation (such as the



If the grass is wet and the sprinkler is on, did it rain? A deep question for causal reasoning

workings of a 'blicket-detector').

All of these views (logical rule learning and information processing, developmental stages and cognitive architectures, connectionist networks and dynamical systems, Bayes nets and structure learning, and others) continue to be influential and active avenues of research. By reviewing them as history I do not mean they are historical, but at this point I want to turn to recent advances in both computational and developmental cognitive science. Just like previous parallel advances, these too have something to say to one another.

1.2 Theories of Theories - Current Developmental Views

What does current experimental research tell us about Turing's vision of the child as an empty notebook? What is the amount of content, what is the language, and how do children go about filling it with new ideas?

We would say, not that children are little scientists but that scientists are big, and relatively slow, children.
– Alison Gopnik and Henry Wellman

Regarding the questions of knowledge representation and acquisition, a powerful set of ideas mentioned in the previous section was that of 'theory theory' and 'the child as scientist' [24, 22, 123, 66, 69, 67, 152, 154, 193, 73]. On this view, children can evaluate and adopt rich structures of knowledge that go far beyond the sparse data they're given, similar to the way a scientist can propose general principles from limited observations and evaluate them. The 'theory theory' posits that the knowledge itself is represented as something like a scientific theory. The 'child as scientist' view adds that the process of acquiring new knowledge is itself science-like, in that children conduct experiments and design interventions [165], search for new data when needed, isolate variables [32], understand when evidence is confounded

[153], are sensitive to how the data was generated [73], and so on.

Regarding the question of 'amount of initial knowledge', the empirical answer uncovered by researchers in child development over the past decades is "Turing's notebook is not empty, but it is not overly cluttered". Researchers in cognitive development have discovered infants and young children understand several abstract principles which are present early on, across cultures, and shared with non-human animals [168, 169, 5, 194, 34, 128, 155, 4, 24, 27]. These principles are organized into systems of core knowledge for specific domains, with infants maintaining qualitatively different expectations for entities classified under different 'core' domains, such as geometry, number, physics, sociology and psychology.

Opinions may vary as to the complexity which is suitable in the child machine.
– Alan Turing

Thus Turing's notebook might actually be several notebooks, filled with chapter headings, outlines and cross-references, even if they do not contain much specific propositional knowledge. The specific focus of this dissertation will be on intuitive physics and intuitive psychology, the ability to reason productively about mechanical objects and goal-directed agents, and so I provide a bit more detail on those below:

Intuitive Physics As early as 2 months and possibly before, infants already posses some notion of object persistence, continuity and cohesion. They expect objects to follow relatively smooth paths, not wink out of existence, and not act at a distance [168]. Infants also do not expect drastic changes to physical properties (although what determines a physical property and whether size, color or shape matter is subject to some debate). Infants have a notion of object solidity [169], expecting objects not to pass through one another. Many of these expectations are limited to 'cohesive' objects, not applying to things such as sand piles. Over the months following birth, infants develop more adult-like intuitions regarding physical objects. They have a notion of gravity, expecting released objects to fall down [110, 124], and slowly develop ideas regarding inertia (e.g. objects should not simply stop

for reason) and support [81] (e.g. the know what configuration prevents objects from falling down). Infants can also predictively look and reach towards moving objects, although they have a more difficult time reaching when these objects go behind occluders [80]. By 5 months, they have already developed different expectations about solid and non-solid objects [80].

Intuitive Psychology There is a wealth of experiments showing that pre-verbal infants attribute agents with goals, morals and efficient planning. Young infants can encode goals, and expect agents to act efficiently to achieve goals, subject to environmental constraints [168, 34, 33]. They distinguish first anti-social agents from neutral agents, and then pro-social agents from both, preferring pro-social agents that help others over neutrals, and neutrals over anti-social agents that hurt or hinder others [94, 75, 76, 74]. There is some debate about how infants categorize agent and non-agents. While perceptual features such as faces or eyes are useful, they are not necessary [85]. Infants are also sensitive to self-propelled motion [146, 132], efficient movement towards goals given possible actions, and social responsiveness [33, 52].

The ideas of core-knowledge and the 'child as scientist' impose several constraints on what a formal account of human development should look like. Both ideas are concerned with theories of how the world works, the hidden underlying causes that produce observations. How well do the computational accounts in the historical review capture these ideas? Connectionist networks, for the most part, are not concerned with building in core knowledge, nor with anything like a theory. Systems of rules have more the structure of a theory, but are perhaps too brittle, and fail to account for learning and changing entire systems of concepts a-la Kuhn [101]. Of the frameworks reviewed, Pearl's Causal Nets come closest to the notion of finding structured hypotheses to explain the data, but they too are constrained and cannot

account for higher-level aspects of a child-like theory[3]. It is equally unclear how a Bayes Net could account for core knowledge principles like 'objects follow smooth paths' and 'agents have goals'.

In the next section, I turn to a computational framework based on recent advances in computational modeling [50, 142]. This framework combines the strengths of the symbolic and statistical traditions into structured probabilistic models that use Bayesian statistical inference. In cognitive science in particular it has led to a better understanding of high-level human cognition [178], and is currently best-suited to rise to the challenges presented by advances in developmental research.

1.3 Hierarchical Bayesian Models over Rich Structures

The following framework is based on the idea that people reason about by the world by considering how hypotheses can account for data. On this proposal, a reasoning system evaluates a hypothesis h about how the world works, by taking into account the observed data d, and some prior assumptions, background knowledge, beliefs and constraints given the domain theory T. A hypothesis about the world can be about the goal of an agent, the existence or shape of an unseen obstacle, the underlying force law of some dynamics, the causal mechanism responsible for a toy working in some way, and so on.

The degree of belief that a rational learner should assign to some hypothesis is

[3]Consider for example a theory of illness that posits diseases as the cause of symptoms. A Causal Nets story might imagine children hypothesizing various different causal nets until they hit upon the right one for particular diseases and symptoms and understanding the specific causal direction, but nowhere in this learning process or final outcome is there the basic theoretical statement "there are two types of things in the world, diseases and symptoms, and diseases cause symptoms" [178].

equivalent to the posterior probability of that hypothesis, calculated using Bayes' rule:

$$P(h|d,T) \propto P(d|h) \cdot P(h|T). \tag{1.1}$$

This equation captures the way beliefs are updated as the result of an interplay between the prior knowledge of an intelligent system (adult, child, machine), and the need to account for the data. The likelihood term $P(d|h)$ assesses how likely the data is given the hypothesis, while the prior probability $P(h|T)$ indicates how 'reasonable' the hypothesis is, independent of the data. Children's mental development can then be seen as a process of theory revision - strong assumptions about how the data was generated can be changed given conflicting data.

Arthur: Camelot!
Patsy: It's only a model!
Arthur: Shh!
(Monty Python and the Holy Grail)

This formal generative approach is expanded by specifying multiple levels of a 'theory hierarchy' (and giving us *Hierarchical* Bayesian Models). Domain theories then constraint models of particular scenarios, and domain theories are in turn constrained by higher and more abstract principles [92].

It is a pretty picture, but it is only a sketch of a general framework, and the rational belief updating mechanism (Eq. 1.1) is only the basic skeleton of inference. The real challenges – the flesh and nerves – are these:

Explain how the world works by specifying the actual theory structure of the hypothesis spaces

Explain how learning works by giving rational, realistic learning algorithms for exploring these spaces

To better understand the first challenge, consider how the HBM formalism might capture the intuitive theory of psychology. The observed data d we want to explain

are series of *Actions* ("Why did John open the box?"), while the unobserved things we use as explanations are mental constructs such as $Beliefs$ and *Goals*. How do we compute $P(Goals, Beliefs|actions)$? Simple, says the Bayesian updating mechanism:

$$P(Goals, Beliefs|Actions) \propto P(Actions|Goals, Beliefs) \cdot P(Goals, Beliefs) \quad (1.2)$$

But how do we get the likelihood of actions given goals and beliefs, or the prior on goals and beliefs? That is the hard part. The 'theory' of agents is that they act efficiently in order to achieve goals. This can be formalized as a rational planning model, the sort of thing developed for economics, robotics and artificial intelligence [133, 9]. Imagine for example a robot with a planning procedure. If the robot is told its goal is to get an apple (high utility for states where it has the apple), and the robot believes the apple is in a box (high probability on states where the apple is in the box), then the robot can use a planning procedure to produce a sequence of actions that will get it to its goal (open the box and get the apple). So, the planning procedure gives us the probability of taking certain actions given goals and beliefs, which is the likelihood we were after.

By assuming that this is how people work, we can explain their actions. If we see someone reaching for a box and grabbing an apple, we can say that they probably like eating apples, and that they believed the apple is in the box. We can incorporate different knowledge into this story, too: if we think John hates apples, we might think John believed there was something else in the box.

Chapter 2 expands on 'intuitive psychology as inverse planning'. The main take-away of the previous paragraphs is that while HBMs (Eq. 1.1) can formalize the idea of children rationally updating theories, they are not the end of the story. One still

has to work hard to specify the right theories, their structure, and their basic units. This is a challenge, but it is a challenge that HBMs and cognitive development can work to solve jointly.

What about the second challenge mentioned earlier, that of learning? This deserves its own section.

1.4 Learning and the Algorithmic Level

Even if we built the right formal theories for each core domain using findings from core knowledge, we still would not be happy [4]. We would still have to explain how learning new theories happens. In this section I lay out common learning-based objections to the HBM formalism, and point in the direction of a solution that will be developed in Chapter 4.

On some level, Eq. 1.1 explains how learning happens: A rational agent should shift probability mass on theories as new data comes in, taking into account both the fit to data and theory simplicity. But on a different level, this is not a satisfying statement. The objections to this 'explanation' usually fall into one of the following inter-related groups:

The Objection of Limited Thought "HBMs are quite successful in capturing some of the reasoning of children and adults, but they only succeed because the hypothesis spaces you pre-defined is small. You can capture children moving from theory A to theory B, by assuming the hypothesis space is limited to A and B and that as data comes in more probability is placed on theory B, but children can also potentially think of C, and D, and an infinite number of

[4] Actually we would be extremely happy. But we wouldn't be *satisfied.*

things that don't go into your hypothesis space at all."

The Objection of Infinite Incredulity "So you can define very large or infinite hypothesis spaces. But your view of learning is then a shifting of probability on a very large or infinite space. How can you honestly suggest children and adults have parallel access to each hypothesis in such spaces? Children probably consider at most only 2 or 3 options at a time."

The Objection of Mad Nativism "If you define the entire space of hypotheses, you're not actually learning anything new, you're just testing and confirming things you already new. This is Mad Dog Nativism. Are you honestly suggesting that the move from Newtonian Physics to General Relativity should be captured by considering all the possible theories of physics, and saying that Einstein shifted probability mass to General Relativity? If General Relativity was already a possible thought to consider, in what meaningful sense did Albert come up with anything new?"

These are reasonable objections and concerns that need to be addressed. The first objection is addressed by allowing for larger hypothesis spaces, but then one runs into the second objection. I said that Eq. 1.1 is the Bayesian learning story 'on some level'. Usually the term 'on some level' is a figure of speech, but in this case it can be made more precise. According to the Marr-Poggio proposal, we need to think about cognitive systems using three different levels of analysis: The computational, the algorithmic, and the implementation [112]. The computational level defines what the task of the system is, the algorithmic level specifies how representations are manipulated to achieve the task, and the implementation level gives the physical realization of the algorithm. There are usually many algorithms that can solve a given task, and many implementations for a given algorithm.

The previous section described HBMs at the computational level: The task of the mind is to reverse-engineer the structure of the world, aided by Bayesian inference. This is the 'level' of Eq. 1.1, and this is the level that the objections are aimed at. But the objections can be (mostly) answered by referring to the algorithmic level. The Objection of Infinite Incredulity scoffs at parallel access to large spaces, but an intelligent machine has no more parallel access to these spaces than children or adults do, and yet computational researchers are able to do inference over such spaces. They do it by using algorithms to implement the inference, algorithms that usually consider only a few hypotheses at a time and are prone to backtracking. Such algorithms only approximate the ideal level, and their dynamics are not that of an ideal rational process. They are rational approximations, motivated by the underlying theory expressed at the computational level.

Therefore, it might be better to equate the learning process of a child or adult with the process of a rational algorithm searching through a space of theories. This suggestion also addresses to some degree the Objection of Mad Nativism. By neglecting the algorithmic level the Objection of Mad Nativism is true, but it is true in an uninteresting way. It is similar to stating that a person that commands the grammar of the English language can never actually say or think anything new, because the grammar defines an infinite space of utterances and sentences that are (in some mathematical sense) 'there'. On some level this is true, but it is an uninteresting level. People can generate sentences by sampling from their grammar (their 'space' of sentences), and they can generate ideas and theories by sampling from their space of thought.

Chapter 4 considers the algorithmic level of learning in more depth. In particular it examines the similarities between a class of algorithms (known as Markov Chain Monte Carlo) and the learning dynamics of children, and relates the algorithmic

process to theories of conceptual change.

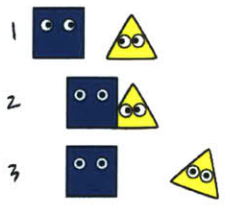
At this point I've presented current views coming out of development, and a general outline of how a computational framework can make contact with them. At this point a reader might raise a more general objection: Suppose these models provide both a conceptual and behavioral fit to the current views of cognitive development – which remains to be shown – what is the alternative? Can a formalism that assumes much less mental machinery account just as well for the data? This challenge appears in all the following chapters, and I address it in general in the next section.

1.5 Cues, Classifiers, Trees and Rules, and Other Things that Probably Won't Work

Can we do without all the mental modeling baggage? There's certainly a long-standing tradition that tries, which I'll refer to as the Classifier-Based approach[5]. Here is a compressed one-sentence summary of the Classifier-Based approach, lumping together several strands of different research:

"Given that children and adults receive input X and produce output Y, find *something* that can take in the properties of X to produce Y. "

The Classifier-Based approach is different from Good Old Fashioned Behaviorism in that the 'output' can be a mental state or percept. Think of a person seeing two googly-eyed shapes colliding. Such a scene can produce in the person the following mental sensations: That the two shapes are 'agents', that the first agent had the 'goal' of crashing into the other one, that this crash 'caused' the other one to move,



[5]This term is somewhat misleading, in that the tradition is broader than just classification. But it's useful to have a label, and Classifier-Cue-Rule-And-Similar-Things-Based approach is a mouthful.

that the first agent is is 'heavier', and so on.

Agency, goal, causality, mass. Behaviorism is loath to consider these mental percepts as targets of research. But the Classifier-Based approach is quite willing to consider them, going back at least to Michotte's studies of the mental percept of causality [119].

The Classifier-Based approach is different from a theory-based approach in that its primary concern is with finding properties of the input to use for an input-output mapping between the perceptual input and the mental output. For example, some of the percepts of the previous example might be captured by the **rule** "If something started from rest, then it is an agent" [145]. Or we might say that faces are a good **cue** for agency. Similarly, we might say that people have a **heuristic** such that if an object's post-collision velocity is greater than its pre-collision velocity, people perceive that object to be lighter [179]. We might posit innate **perceptual analyzers** that trigger the sensation of causality when the motion of one shape is followed by the motion of a second shape without a spatio-temporal gap [119]. We could also build **decision trees** that chain together a bunch of yes-no questions about the properties of the scene to produce the mental output [3].

These proposals ignore (or deny) any underlying theory connecting the input and output. We don't need an understanding of how causality works to classify a scene as belonging to an instance of 'A caused B to move'. We don't need a theory of agency – with its goals, beliefs, plans, intentions – to classify A as 'a bad guy'. Once this is accepted, the task of research is then to uncover the relevant features and cues for any given situation (e.g. do velocity and angle play a role in mass judgment, or just velocity?), and to mark the borders of the perceptual analyzers and rules (e.g. at what point do spatio-temporal gaps nullify the feeling of causality in collision events?).

What then of development? The Classifier-Based approach is concerned with finding the long list of various innate cues, classifiers, heuristics and analyzers present from birth. Some avenues of research then suggest how new rules and decision-nodes can be acquired, accounting for shifts in infant judgments as they grow older [159, 3].

The approach is supported by experimental evidence and methodological simplicity. Many mental percepts appear fast, automatic, immune to experience and present from birth (Chapters 2, 3 and 5 discuss this in more detail). Also, if the Classifier-Based approach can explain a mental judgment just as well as a theory-based one, it should be preferred because it posits fewer entities.[6]. And yet this approach is incomplete.

I don't mean to deny the reality of cues, classifiers, heuristics and so on. People from infancy onwards do seem to have special fast detectors on the lookout for aspects of physics and psychology (or "mechanical and social causality" [143, 146]). But they cannot be the whole story. This statement is explored in the rest of the thesis, and the arguments are broadly these:

1. Our intuitive knowledge can reckon with an infinite number of questions, contingencies and scenarios, but any new question might require a new feature or cue or rule. The 'simplicity' of the Classifier-Based approach collapses under the sheer number of features to consider. For example, we need separate cues to answer how a tower will fall, in what direction it will fall, and how it will

[6]For example, suppose the two approaches try to explain how young children know an agent moving towards a location has that location as its goal. The theory-based approach might posit that young children have a mental model of what agents are like: Agents can plan to achieve goals, they have some belief about where the goal is, and they take efficient series of actions to get to their goals. The Classifier-Based approach might counter with 'IF a shape started from rest and is moving towards an object, THEN that object is its goal', or 'the shrinking distance between an agent and an object is a cue that the object is the agent's goal', or simpler still 'the shrinking distance is a cue that can predict in the future the shape will move towards that object'.

scatter , while a single theory-based model can answer all of these and a large number of other questions as well [12].

2. The same cue or feature or rule can lead to different mental states, and different cues could lead to the same state, depending on the situation. For example, moving someone in the direction of their previous motion seems a useful cue for 'helpfulness', but that action could result in a person being pushed off a cliff. Similarly, being 'helpful' might sometimes require us to move towards someone, and sometimes further away.

3. We don't yet have a formalization of core knowledge, but its principles are not stated in anything like a cue-based form [168]. The idea that 'agents act efficiently to achieve goals' is a proto-theory of how agents work, and a rule such as "IF something acts efficiently to achieve goals THEN it is an agent" is simply begging the question. Similarly, the idea that 'objects should follow smooth paths and maintain cohesion' is a proto-theory of how physical objects work, not a statement about the right cues for detecting physical causality.

Classifiers are real, and important. They might be fast ways of focusing on a small part of a hypothesis space. But they don't replace hypotheses of how the world works.

Chapter 2

Help or Hinder*

What is hateful to you, do not do to your fellow. This is the whole Torah, and the rest is commentary, go and learn. — Rabbi Hillel the Elder, Talmud Bavli

2.1 Introduction

Suppose a person suddenly finds herself on board the ship of Odysseus, just as it draws near the island of the Sirens. Unaware of the Greek classics, she watches in horror as Odysseus is bound hand and foot to the ship's mast with tight ropes, hears him yelling and begging to be set free. Rather than listening to their king, the men add more cords and draw the ropes tighter. This person would probably think

The Sirens, stamnos vase, 480-470 BCE

*Joint work with Owen Macindoe, Chris Baker, Owain Evans, Noah Goodman, and Josh Tenenbaum

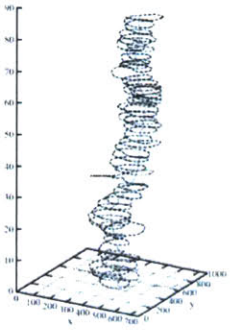
Odysseus' sailors are sadistic brutes. But how quickly she would change her mind if she knew the disastrous fate of those lured by the Sirens' call [103].

While this example is fanciful, people constantly encounter similar situations - situations requiring them to think about the social intentions driving the actions of their peers, their friends and enemies. As a more prosaic example, consider a child whose mother just slapped her wrist after she reached for a hot stove. What should the child make of the situation? Is the mother intending to hurt, or warn? The child might reasonably expect the mother is trying to help her, much like in the past, and so reaching for a hot stove is dangerous. Compare this to a case where an older sibling just hit the child, and instead of a hot stove the child reached for a shiny new toy. In this case the child would probably realize something about the preferences of her brother, rather than conclude that the new toy is dangerous.

Social inferences are fast, intuitive and robust. They happen automatically, with people reading social meaning into even extremely impoverished visual displays: A short video of bland geometric shapes moving in a 2D world causes adults to spontaneously attribute to these shapes a host of aims and intentions [79]. Some of the attributed goals are simple, like reaching an object or a location. But people also attribute complex social goals, such as helping, hindering or protecting another agent. Recent studies suggest that not only adults, but also pre-verbal infants make complex social goal attributions when looking at simple displays of moving shapes [143, 100, 75], or watching puppets interact [76, 74]. Reasoning about social behavior thus seems early-emerging and universal, and was even suggested as a candidate core knowledge system [168].

How do people make these inferences? What is the structure of knowledge that accounts for this kind of understanding? Is this knowledge even structured in any high-level sense? One approach sees this knowledge as emerging from the physical

and perceptual cues of the observed stimuli. On this view, the visual system automatically uses perceptual cues to reconstruct the social nature of objects and scenes, just as it reconstructs their three-dimensional nature [147]. Advocates of this approach point out the rapidity and robustness of goal attribution, arguing that these require an 'automatic' inference built on visual perception, without the need for mediation from higher cognition. This "Cue Based" approach is present in computer science and machine learning, as well as psychology and neuroscience. In computer science, some researchers have focused on identifying useful features in the visual scene that will allow them to automatically categorize different motions into conceptual categories. In psychology, this approach goes back at least to the work of Michotte [119], who extensively varied many perceptual cues to examine their effect on 'higher-level percepts' such as causality and animacy. More recently in cognitive science, this viewpoint has been developed by researchers such as [148] and by [16].



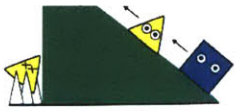
Small 'relative vorticity', a good cue for courting behavior? Adapted from Barrett et al. (2005).

It is easy to see how low-level perceptual cues might explain some simple object or location-directed goals. For example, the shrinking distance between an agent and an object is a good cue for inferring which object is the agent's goal. Beyond location-based goals, this approach was also used to explain simple agent-directed behavior such as chasing and fleeing [48]. Building on these successes, adherents of the Cue-based viewpoint could argue that any goal inference can in principle be captured using perceptual cues, if only the *right* perceptual cues could be identified [11]. However, further consideration shows that in the case of social goals – and abstract goals in general – such a "Cue Based" account becomes problematic. Social goal inference is challenging because actions in and of themselves do not appear to to hold intrinsic moral and social content (as pointed out by philosophers such as Hume[1] [82]). A particular action can not be morally or socially evaluated based

[1] "Take any action allow'd to be vicious: Wilful murder, for instance. Examine it in all lights,

purely on its observable physical description. Rather, the social evaluation of actions stems from the mental motivations assigned to the acting agents, mental motivations which are unobservable and need to be inferred. More explicitly, the perceptual cue approach does not easily account for the fact that the same actions could be interpreted completely differently - moving towards someone could be seen as helpful or harmful, depending on the unobserved goal of the agent. Even hitting someone, as in the case of the child reaching for a hot stove or a shiny toy, can be seen as helping or harming that person depending on the situation and the intentions of those involved.

To examine the possible difficulties with the 'Cue-based' approach more concretely, consider the study described in [75], in which infants see a two-dimensional agent (say, a yellow triangle with eyes) placed at the bottom of the hill. The agent then moves up the hill, but fails to reach the top. During one of the attempts, another agent (e.g. a blue square) enters the scene and either moves the triangle up the hill, or moves it down the hill. Based on these scenes, infants make predictions and show preferences which suggest they understood the square was 'helping' or 'hindering' the triangle. The "Cue-based" account might argue that making this inference is merely a case of using the right motion features. For example, infants may judge the motion of the square as helping simply because the square is moving the triangle in the direction the triangle was last observed moving on its own. However, consider that pushing the triangle down could be helpful if there had been previous evidence



Helping?

and see if you can find that matter of fact...which you call vice. In which-ever way you take it, you find only certain passions, motives, volitions and thoughts. There is no other matter of fact in the case. The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflexion into your own breast, and find a sentiment of disapprobation, which arises in you, towards this action...It lies in yourself, not in the object. So that when you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it. Vice and virtue, therefore, may be compar'd to sounds, colours, heat and cold, which, according to modern philosophy, are not qualities in objects, but perceptions in the mind" (Hume, A Treatise on Human Nature)

that there is something dangerous at the top of the hill, or that the triangle's goal is at the bottom of the hill.

In fig. 2-1, I show several examples of actions involving two agents in pursuit of goals, using a maze-like version of the Hamlin et al. task. The larger agent in this case is able to push a boulder around, and cannot be moved by the small agent. This similar to how the second agent in the Hamlin task has more affordances than the first agent, that cannot climb the hill on its own. The larger agent pushing a boulder out of the smaller agent's path could be seen as a helpful action, allowing the small agent to reach its goal on the other side of the boulder. However, this same action could also be seen as selfish, if the large agent merely pushed the boulder out of the way in order to get some reward on the other side for itself. A particular action - such as moving towards or away from the other agent, pushing it or moving objects - could be interpreted as helping, hindering or selfish actions depending on the context. The 'Cue-based' approach does not easily account for this, nor for the fact that completely opposite actions could be interpreted as belonging to the same higher-level goal. For example, if the goal of the large agent is to help the small agent, in one situation it might require getting closer in order to push it along, in other situations it might require moving further away to not block the passage. Even in such elementary cases the apparent simplicity of the Cue-based account fades away, requiring more and more cues as the number of possible scenarios grows larger.

In contrast to such a perceptual-cue based account, I propose that the complexity and robustness of social goal inference require structured models which can incorporate rich abstract knowledge. More specifically, I suggest social goal inferences can be captured by a generative Bayesian formalism that explicitly uses the notions of state, agent, and utility. This view is part of a general formalism in cognitive science and AI which involves specifying the underlying processes that generate potentially

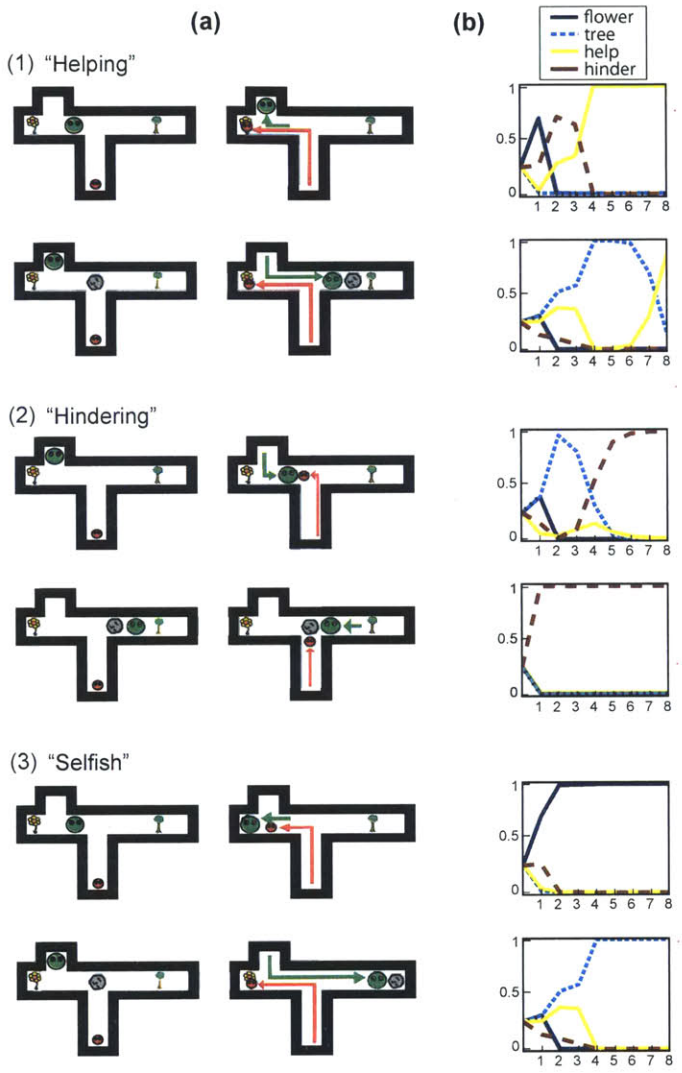


Figure 2-1: 6 Examples of social interactions between agents, and the model inferences made on their basis. **(a)** The examples show 2 snippets each of "helpful", "hindering" and "selfish" behavior on the large agent's part. The left panel shows the starting positions of the agents, the right panel shows the end position. Colored arrows indicate the sequence of movement. **(b)** The posterior probability of the large agent's goals as the scenario unfolds, according to the Inverse Planning model.

observable data, and then reasoning back from the actually observed data to the hidden underlying causes. The formalism has proven useful in understanding cognitive domains such as perception [197] and motor control [99], and recently gained popularity in areas of higher-level cognition such as causal reasoning, object properties and relations, category and classification, intuitive physics and general knowledge acquisition [178, 12, 61, 195]. I will briefly review how this formalism was applied in the domain of goal inference, then argue that an extension of this framework to social goals can succeed where Cue-based methods cannot.

In the domain of action understanding and goal inference, the key underlying process generating the data is planning. Planning takes an agent from goal representations and beliefs to observable actions. As the inference of goals and beliefs requires inverting this process – using Bayes' rule to reason back from actions to hidden states – it is referred to as 'Inverse Planning'. Baker et al. [8] showed how one can use Inverse Planning to infer simple goals such as being in a particular location, demonstrating strong correlations between this model and human responses on tasks similar to those used by [79] and[75].

Baker et al. relied on the **The Principle of Rational Action** to describe how a rational agent should act in a given environment. In psychological terms, this principle determines that "An agent will take means to achieve its goals, given its beliefs and the environment it is in". Models of rational action assume agents use rational planning to guide their actions given certain goals and constraints. Such planning models have been developed by economists to explain group and individual behavior, by psychologists and cognitive neuroscientists to explain mental and physical processes in the brain, and by computer scientists in order to build intelligent systems capable of achieving certain aims and goals [133, 150, 189]. The underlying psychological principles can be captured computationally, by considering the

decision-making processes of agents trying to maximize their utility given a possible state of affairs. In this chapter we consider an extension of utility-based planning known as Markov Decision Processes (MDPs). We will describe MDPs and their relevance to psychology and action understanding in the next section.

While the principle of rational action (phrased in terms of utility theory) tells an agent how to act, and while it is possible that social goals can be represented in utility-theoretic terms, the principle of rational action alone does not specify how social goals are represented, what they mean, or how an agent should pursue them. To see informally why such a representation is necessary and useful, consider a Martian that has no idea how to act in human society. The Martian is a rational planning agent, acting by the principle of rationality and capable of planning actions given certain goals. One can imagine giving this Martian an infinite list of simple goals prescribing exactly how to behave in any given situation ("If a friend is thirsty and wants to drink water, give them water, if they want soda, give them soda", etc.). Such an exhaustive list might be technically possible, but would be impractical, unwieldy, and brittle. Instead, we might offer the Martian a general 'Golden Rule': You should act towards others as you wish them to act towards you. This standard of behavior and morality, one form of which is the epigram of this chapter, comes up in many religious and philosophical texts throughout history, from ancient Egypt through the verses of the Mahabharata, from the sermons of Jesus and up to modern times. The idea is simple, yet abstract, and it has wide-ranging implications when combined with rational inference.

"Do unto others twenty-five percent better than you expect them to do unto you, to correct for subjective error"
– Linus Pauling

We therefore propose in this chapter an additional principle, which like the principle of rational action is simple, abstract, and can potentially reduce a great amount of complexity when combined with rational inference - the **Principle of Sympathy/Antipathy**. This principle specifies the representation of social goals, which

can then be combined with the principle of rational action to reason about agents that use rational social planning. Put formally, the positive part of this principle is "In order to help someone, adopt their goal state as your own goal". In more utility-based terms, we define agent A as trying to help agent B if agent A explicitly defines its utility function to depend in a positive way on agent B's utility function:

$$Helps(A, B) \rightarrow U_A(S) = f(U_B(S)) \tag{2.1}$$

Where U_i is the utility of agent i, S is a state of the world, and $f(x)$ is an increasing function of x. So, whatever is good for agent B will be good for agent A, and whatever is bad for B will be bad for A.

Equally important, the negative part of this rule shows us how to go about hurting and hindering our fellows, those that we have antipathy towards. We define agent A as trying to hinder agent B if agent A explicitly defines its utility function to depend in a negative way on agent B's utility function:

$$Hinders(A, B) \rightarrow U(A) = g(U(B)) \tag{2.2}$$

Where $g(x)$ is a decreasing function of x. Now, whatever is bad for B will be good for A, and whatever is good for B will be bad for A. We will later refer to this formalization of the terms 'helping' and 'hindering' as the Principle of Sympathy.

In previous studies in the Inverse Planning tradition [10, 9, 8] the utility of agents was a function of the state of the world. Here we extend utilities to include functions that take in *other utilities*. This makes helping and hindering into a more abstract relationship. Fig. 2-2 shows schematically the move from solitary rational agents to social rational agents. For both agent types, the *environment* produces certain *beliefs*, which combined with the *desires* of the agent dictate its *actions* through

the Principle of Rational Action. Beyond this, social agents now take into account the planning models of other agents they are interacting with, and their desires are defined on the desires of other agents, according to the Principle of Sympathy: the utility function of a social agent depends to some degree on the utility function of other agents. This relation between utilities is an abstract relationship, which is world independent and extends across a multitude of different scenarios. Given a new world with new sets of actions, a helpful agent could take new actions while maintaining the same relationship with the target agent.

Our challenge in this work is to show that the computational model provides a qualitative and quantitative fit to rapid human social goal inferences, and can capture fast human judgments from impoverished and unfamiliar stimuli. To do this we use stimuli in the form of dynamic visual displays, showing agents moving about in simple 2D maze-worlds. This paradigm was chosen to resemble previous stimuli used in many studies with children and adults. These stimuli allow us to compare quantitative and qualitative performance between human performance and computational models. We also compare our results to those of an alternative cue-based model which makes inferences directly from visual cues such as distances between agents or distances between goals. With this comparison we show our approach can capture social goal inference across different scenarios in a way that resembles human inference, and which the cue-based alternative in its current form cannot account for.

2.2 Computational Framework

Our framework assumes that people represent the causal role of an agent's goals in terms of an intuitive principle of rationality[35]: the assumption that agents will tend to take efficient actions to achieve their goals, given their beliefs about the world.

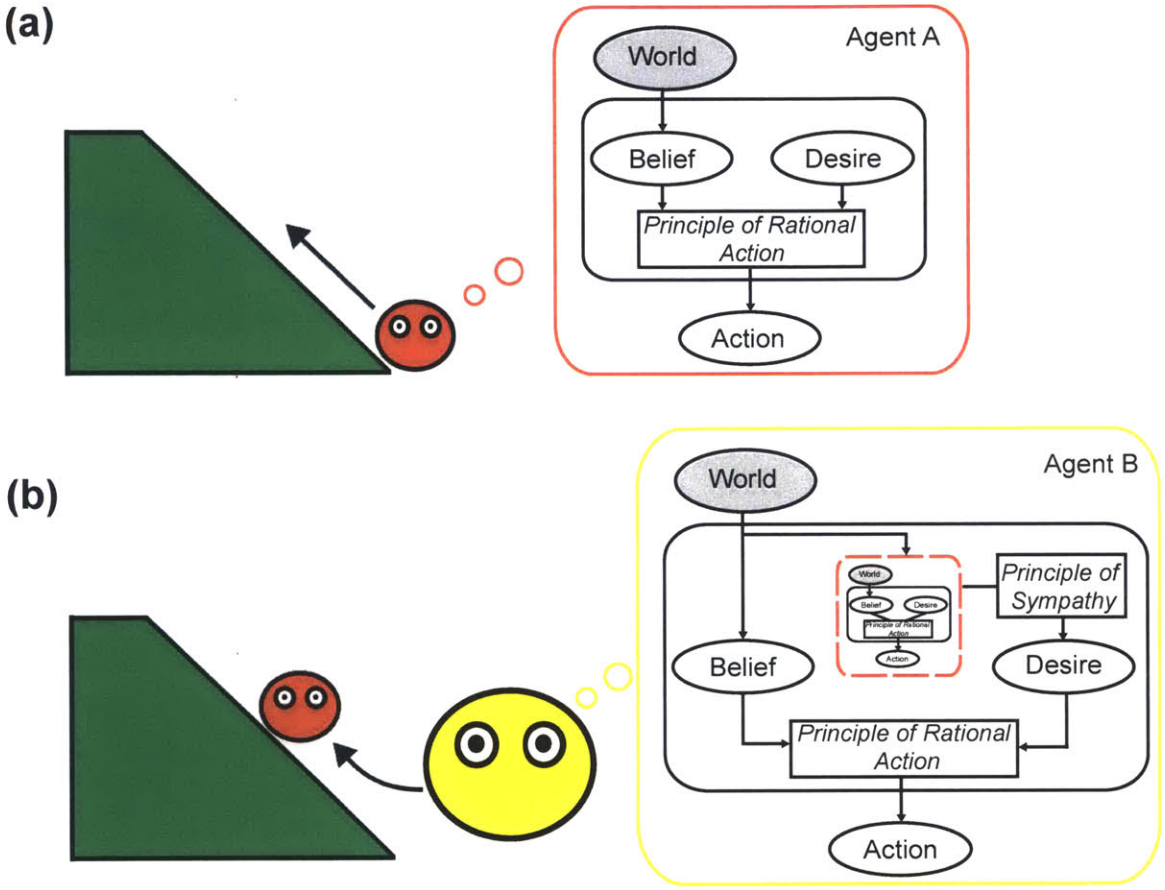


Figure 2-2: Theory of Mind and the Principle of Rationality, with extension to multiple agents and social goals. **(a)** A model of a simple agent with beliefs about the environment formed from experience with the world, and certain desires (such as getting to the top of the hill). The agent chooses the appropriate next step (moving up the hill), assuming a principle of rationality dictates its planning. **(b)** The extension to multiple agents with social goals. The social agent constructs a model of the other agent, from observing its actions in the world. The desires of the social agent are dependent on the other agent through the principle of sympathy, so that if the large agent wants to help the small agent, and believes that the small agent wants to move uphill, then the large agent will push the small agent uphill.

The principle of rationality can be formalized using different planning procedures. One such successful planning procedure which explicitly uses the notions of agent,

state and goal is probabilistic planning in Markov decision problems (MDPs). Previous work has successfully applied Inverse Planning in MDPs to explain human inferences about the object-directed goals of maze-world agents[8].

Multiagent extensions of MDP-based Inverse Planning were considered by [10], capturing simple relational goals between agents such as chasing and fleeing. In this work, we use similar multiagent MDPs to formally present a framework for modeling inferences of more complex social goals, such as helping and hindering, where an agent's goals depend on the goals of other agents.

The structure of this section is as follows: We begin by describing the 'generative/forward' direction of planning in a multiagent MDP, giving a mathematical formulation. We then describe the structure of the reward functions the agents have, distinguishing between object-directed reward and social rewards. We distinguish between simple, non-social agents and complex, social agents, based on their reward function and their own planning model of other agents. Finally, we describe the Bayesian inversion of the multiagent planning process which leads us from observed actions to the joint inference of object-directed and social goals. We will use stimuli similar to that used in the experiments to give concrete examples of the notions detailed here.

2.2.1 Planning in multiagent MDPs

An MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$ is a tuple that defines a model of an agent's planning process. $\mathcal{S}$ is an encoding of the world into a finite set of mutually exclusive *states*, which specifies the set of possible configurations of all agents and objects. $\mathcal{A}$ is the set of actions, and $\mathcal{T}$ is the transition function, which encodes the physical laws of the world, *i.e.* $\mathcal{T}(S_{t+1}, S_t, A_t) = P(S_{t+1}|S_t, A_t)$ is the marginal distribution over the

next state, given the current state and the agent's action (marginalizing over all other agents' actions). $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, which provides the agent with real-valued rewards for each state-action pair. γ is the discount factor, which dictates how much future rewards diminish in their value compared to the immediate reward. To make this more concrete, consider the simple maze-world presented in Fig. 2-1.

The set of possible actions $\mathcal{A}$ for each agent is *(move up, move down, move left, move right)*. The states $\mathcal{S}$ would be a set of 2-dimensional grid-coordinates of the agents - $((x_{large}, y_{large}), (x_{small}, y_{small}))$. Assuming we use an 8-by-5 grid to specify the location of the agents, the initial state of the agents in **a.1** is $((3,4),(4,1))$. Consider that the large agent now attempts to take the action *move left*, and the small agent takes the action *move up* . If the actions aren't noisy, the transition function would place a probability of 1.0 on the next state being $((2,4),(4,2))$.

The following subsections will describe how $\mathcal{R}$ depends on the agent's goal G (object-directed or social), and how $\mathcal{T}$ depends on the agent's type (simple or complex). We will then describe how agents plan over multiagent MDPs.

Reward functions

Object-directed rewards The reward function induced by an object-directed goal G is straightforward. An agent planning under this reward function will take actions to minimize the distance between it and the goal object or goal location, contingent on action costs and environmental constraints. We assume that the reward $\mathcal{R}$ is an additive function of state rewards and action costs, such that $\mathcal{R}(S,A) = r(S) - c(A)$, where $r(s)$ is the reward for being in state s, and $c(a)$ is the cost of taking action a. Basic intuition dictates that an agent with a certain object as its

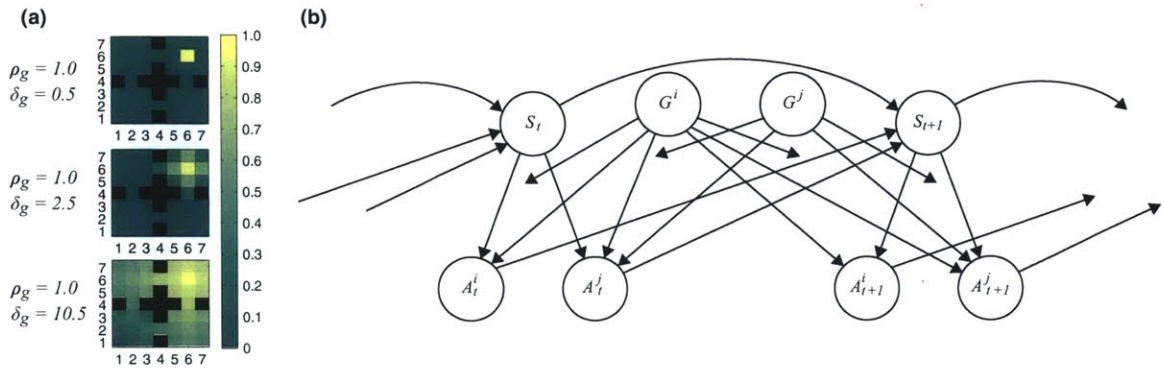


Figure 2-3: **(a)** Illustration of the state reward functions from the family defined by the parameters ρ_g and δ_g. The agent's goal is at (6,6), where the state reward is equal to ρ_g. The state reward functions range from a unit reward in the goal location (row 1) to a field of reward that extends to every location in the grid (row 3). **(b)** Bayes net generated by multiagent planning. In this figure, we assume that there are two agents, i and j, with i simple and j complex. The parameters $\{\rho^i_g, \delta^i_g, \rho^i_o, \rho^j_g, \delta^j_g\}$ and β are omitted from the graphical model for readability.

goal should receive some reward for being in the object-possessing state. However, it is possible to imagine that different objects can induce different rewards in space. Some objects might be rewarding only if one posses them directly - For instance, on a hot summer day in the park, a drinking fountain is only rewarding when one is standing directly next to it. Other objects or locations might be more rewarding the closer one is to them, but still rewarding even if one does not inhabit the goal location itself - One might covet some preferred movie seat, but nearby movie-seats would do fine too. To capture this range, we consider a two-parameter family of reward functions, parameterized by ρ_g and δ_g. These parameters determine the scale and shape of the reward $r(S)$ that one receives for being in a certain state in the following way: $r^i(S) = \max(\rho_g(1 - \text{distance}(S, i, G)/\delta_g), 0)$, where distance$(S, i, G)$ is the geodesic distance between agent i and the goal. By adjusting these parameters, we can go from a 'point-reward' (receiving reward only for being in a certain state) to

a 'diffuse reward field' (receiving more and more reward as one approaches a certain state, up to some maximum value at that state), and from a strong reward signal to a weak reward signal. To see this, consider that with $\delta_g < 1$, the reward function has a unit value of $r(S) = \rho_g$ when the agent and object goal are in the same location, *i.e.* when distance$(S, i, G) = 0$, and for all others the locations the reward is $r(S) = 0$ otherwise (see Fig. 2-3(a), row 1). When $\delta_g \geq 1$, there is a "field" of positive reward around the goal, with a slope of $-\rho_g/\delta_g$ (see Fig. 2-3(a), rows 2 and 3). The state reward has a maximal value of $r(S) = \rho_g$ when distance$(S, i, G) = 0$ (*i.e.* when the agent and the goal object are in the same location). This reward then decreases linearly with the agent's geodesic distance from the goal, reaching a minimum of $r(S) = 0$ when distance$(S, i, G) \geq \delta_g$.

Social rewards for helping and hindering Our formal characterization of helping and hindering goals captures a simple intuition. Suppose a parent moves a child in reach of her favorite toy. Typically, the parent acts not on a selfish desire to move the child out of the way, but on a general desire that the child be in states that are good or desirable (in this case, able to reach her favorite toy). The parent can thus be modeled as having a general goal of doing whatever is best for the child. In a narrow setting in which the child has a single goal (e.g. to reach a particular location) the parent is modeled as sharing the child's goal. Within our utility-based framework, we formalize this idea by supposing that the reward function for an agent with a social goal depends on the reward function of a simple agent. More precisely, if A has the goal of helping B, then A's reward is a strictly increasing function of B's reward. If A hinders B, then A's reward is a decreasing function of B's reward. In both cases, A's reward will also depend on the actions A takes. So A still has a purely selfish concern with avoiding costly actions independently (e.g. moving large

And children, with the prattle and the kiss / Soon broke the parents' haughty temper down
– Lucretius, On the Nature of Things

distances). One can imagine cases in which A has self-directed goals on top of the socially defined goals and has to balance between the two (e.g. to get some coffee, but also to move the child towards its favorite toy). In order to keep the distinction between social and non-social clear, we do not consider such examples here, but the extension is quite simple, and was considered in [84].

We now define these social goals in formal notation. For complex agent j, the state reward function induced by a social goal G^j depends on the cost of j's action A^j, as well as the reward function $\mathcal{R}^i$ of the agent that j wants to help or hinder. Specifically, j's reward function is the difference of the expectation of i's reward function and j's action cost function, such that $\mathcal{R}^j(S, A^j) = \rho_o \mathbb{E}_{A^i}[\mathcal{R}^i(S, A^i)] - c(S, A^j)$. ρ_o is the social agent's scaling of the expected reward of state S for agent i, which determines how much j "cares" about i relative to its own costs. For helping agents, $\rho_o > 0$, and for hindering agents, $\rho_o < 0$.

This formal definition captures the intuitive sense of "helping" or "hindering", which does not depend directly on action. For example, in some cases helping requires moving away from an agent, and in other cases moving towards it. The specific action will depend on the specific situation, but they stem from the same abstract relationship between goals. Recall that in Fig. 2-1 we showed some simple examples of possible interactions between social and non-social agents, and how different actions could give rise to the same social goal inference.

Notice that in order for the an agent j to compute $\mathbb{E}_{A^i}[\mathcal{R}^i(S, A^i)]$ it must itself represent agent i as having a planning model by which it chooses its actions. This is a formal requirement of the model, but it also makes intuitive sense - you cannot adjust yourself to the future actions of some agent without any sense of what these actions will be. Even modeling another agent as taking actions randomly is still modeling it as having some kind of planning process (albeit a poor one). we describe

the different agents' planning process below.

State-transition functions

In our interactive setting, $\mathcal{T}^i$ depends not just on i's action, but on all other agents' actions as well. Agent i is assumed to compute $\mathcal{T}^i(S_{t+1}, S_t, A_t^i)$ by marginalizing over A_t^j for all $j \neq i$:

$$\mathcal{T}^i(S_{t+1}, S_t, A_t^i) = P(S_{t+1}|S_t, A_t^i) = \sum_{A^{j \neq i}} P(S_{t+1}|S_t, A_t^{1:n}) \prod_{j \neq i} P(A_j|S_t, G^j),$$

where n is the number of agents. This computation requires that an agent have a model of all other agents, whether simple or complex.

Simple agents We assume that the simple agents model other agents as randomly selecting actions in proportion to the softmax of their expected cost, *i.e.* for agent j, $P(A^j|S) \propto \exp(\beta \cdot c(S, A^j))$.

Complex agents We assume that the social agent j uses its model of other agents' planning process to compute $P(A^i|S, G^i)$, for $i \neq j$, allowing for accurate prediction of other agents' actions. That is, the complex agents model other agents as choosing their actions in rational pursuit of their goals. The next subsection describes the mechanism for multiagent planning.

We assume all agents have access to the correct transition function, which describes the physical dynamics of the world. This is a simplification of a more realistic framework in which agents have only partial or false knowledge about the environment. We also assume that complex agents have access to the correct goals of the agents they are modeling. This too is a simplification which cannot capture, for ex-

ample, an observer modeling another agent as having a false belief over goals (leading to scenarios such as "The triangle thinks it is helping, but actually the circle does not want to go up the hill at all"). For the questions we examine in this chapter, however, such a simpler framework is entirely adequate and allows us to focus on the question of social relations and goals. We return to this assumption in the discussion section.

Multiagent planning

Given the variables of MDP M, we can compute the optimal state-action value function $\mathcal{Q}^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, which determines the expected infinite-horizon reward of taking an action in each state. We assume that agents have softmax-optimal policies, such that $P(A|S,G) \propto \exp(\beta \mathcal{Q}^*(S,A))$, allowing occasional deviations from the optimal action depending on the parameter β, which determines agents' level of determinism (higher β implies higher determinism, or less randomness).

In a multiagent setting, joint value functions can be optimized recursively, with one agent representing the value function of the other, and the other representing the representation of the first, and so on to an arbitrarily high order [196]. Here, we restrict ourselves to the first level of this reasoning hierarchy. That is, an agent A can at most represent an agent B's reasoning about A's goals and actions, but not a deeper recursion in which B reasons about A reasoning about B.

2.2.2 Inverse Planning in multiagent MDPs

Once we have computed the 'forward process' for each agent - that is, the probability distribution over the actions each agent should take in its environment using mutliagent planning - we use Bayes' rule to infer the driving goals of the agents' plans

from their f actions. Put most generally:

$$P(G|A,E) \propto P(A|G,E)P(G|E) \tag{2.3}$$

Meaning that after we observe some set of actions A by the agents in environment E, the posterior level of belief that we assign to an agent having a certain goal G, is proportional to the likelihood of the agents taking actions A (given by the MDP forward planning algorithm) and the prior probability of that goal. We can later compare this posterior distribution over goals to participants' goal judgment.

In order to complete this computation, however, we need to consider the parameters of the agents' reward function. As described in the Reward Functions subsection, the rewards of the agents are parametrized by the 'type' of reward (point, field, etc) as well as by how much a social agent values the reward given to its target agent. Since we do not have a strong sense of what prior knowledge people have regarding these parameters, we assume a uniform prior distribution over a plausible range and integrate them out per scenario. This allows us to capture the best possible combinations of reward functions and goals for different scenarios without committing explicitly to the prior knowledge people might have.

Fig. 2-3(b) shows the structure of the Bayes net generated by multiagent planning, and over which goal inferences are performed.

Put formally, we begin by computing $P(A^i|S,G^i)$ for agents 1 through n using multiagent planning. We let $\theta = \{\rho_g^i, \delta_g^i, \rho_o^i\}^{1:n}$ be a vector of the parameters of the agents' reward functions. We then compute the joint posterior marginal of agent i's goal G^i and θ, given the observed state-sequence $S_{1:T}$ and the action-sequences

$A^{1:n}_{1:T-1}$ of agents $1{:}n$ using Bayes' rule:

$$P(G^i, \theta | S_{1:T}, A^{1:n}_{1:T-1}, \beta) \propto \sum_{G^{j \neq i}} P(A^{1:n}_{1:T-1} | S_{1:T}, G^{1:n}, \theta, \beta) P(G^{1:n}) P(\theta). \tag{2.4}$$

Ultimately, we need to obtain a posterior distribution over goals, not a joint distribution over goals and reward parameters. To generate goal inferences for our experimental stimuli to compare with people's judgments, we integrate Eq. 2.4 over a range of θ values for each stimulus trial:

$$P(G^i | S_{1:T}, A^{1:n}_{1:T-1}, \beta) = \sum_{\theta} P(G^i, \theta | S_{1:T}, A^{1:n}_{1:T-1}, \beta). \tag{2.5}$$

This integration step allows our models to infer the combination of goals and reward functions that best explains the agents' behavior for each stimulus. It also means we do not use unnecessary extra parameters to fit participant behavior.

Before moving to the experiment, consider this model's behavior on the simple scenarios shown in Fig. 2-1. For example, in **(b)**, the first case of 'hindering', the large agent begins by moving down. This behavior is mainly consistent with the large agent having the goal of getting to the flower or tree, and so the model places more probability on these goals. As the agent moves right, the model reasons that such behavior is inconsistent with having the flower as a goal, for if this was the case a rational planner should probably move left. Appropriately, the model places most of its certainty on the large agent having the tree as a goal. However, on the next steps the large agent stays put, blocking the small agent on its way to the flower. The model rapidly infers that this is actually hindering behavior, and maintains that inference until the end of the scenario. The other example of hindering is more

clear-cut than this, and the model correctly and quickly matches this intuition.

2.3 Experiment

We designed an experiment to test the Inverse Planning model of social goal attributions in a simple 2D maze-world domain, inspired by the stimuli of many previous studies involving children and adults[79, 51, 180, 48, 100, 75, 149]. We created a set of videos which depicted agents interacting in a maze. Each video contained one "simple agent" and one "complex agent", as described in the Computational Framework section. Participants were asked to attribute goals to the agents after viewing brief snippets of these videos. Many of the snippets showed agent behavior consistent with more than one hypothesis about the agents' goals. Data from participants was compared to the predictions of the Inverse Planning model and a model based on simple visual cues that we describe in the Modeling subsection below.

2.3.1 Participants

Participants were 20 adults from the MIT subject pool, 8 female and 12 male. Mean age was 31 years.

2.3.2 Stimuli

We constructed 24 short animation sequences ("scenarios") in which two agents moved around a 2D maze, shown in Fig. 2-4. The maze always had the same layout and always contained two potential object goals (a flower and a tree). In 12 of the 24 scenarios the maze also contained a movable obstacle, a boulder, to increase the number of possible ways in which the large agent could interact with a small agent.

The scenarios were designed to satisfy two criteria. First, scenarios were to have agents acting in ways that were consistent with more than one hypothesis concerning their goals, with these ambiguities between goals sometimes being resolved as the scenario developed (see Fig. 2-4(a)). This criterion was included to test our model's predictions based on ambiguous action sequences. Second, scenarios were to involve a variety of perceptually distinct plans of action that might be interpreted as issuing from helping or hindering goals. For example, one agent pushing another toward an object goal, removing an obstacle from the other agent's path, and moving aside for the other agent (all of which featured in our scenarios) could all be interpreted as helping. This criterion was included to test our formalization of social goals as based on an abstract relation between reward functions. On our formalization, social agents act to maximize or minimize the reward of the other agent, and the precise manner in which they do so will vary depending on the structure of the environment and their initial positions.

The agents in the stimuli were represented as colorful circles with large eyes, similar to those depicted in [75]. Each scenario featured two different agents, which we call "Small" and "Large". Large agents were visually bigger and are able to shift both movable obstacles and Small agents by moving directly into them. Large agents never fail in their actions, e.g. when they try to move left, they indeed move left. Small agents were visually smaller, and could not shift agents or boulders. In our scenarios, the actions of Small agents failed with a probability of about 0.4. Large agents correspond to the "complex agents" introduced in Section 2, in that they could have either object-directed goals or social goals (helping or hindering the Small agent). Small agents correspond to "simple agents" and could have only object goals. The "action" of an agent was depicted by it squeezing in the direction in which it was attempting to move, and if the action was successful the agent moved into

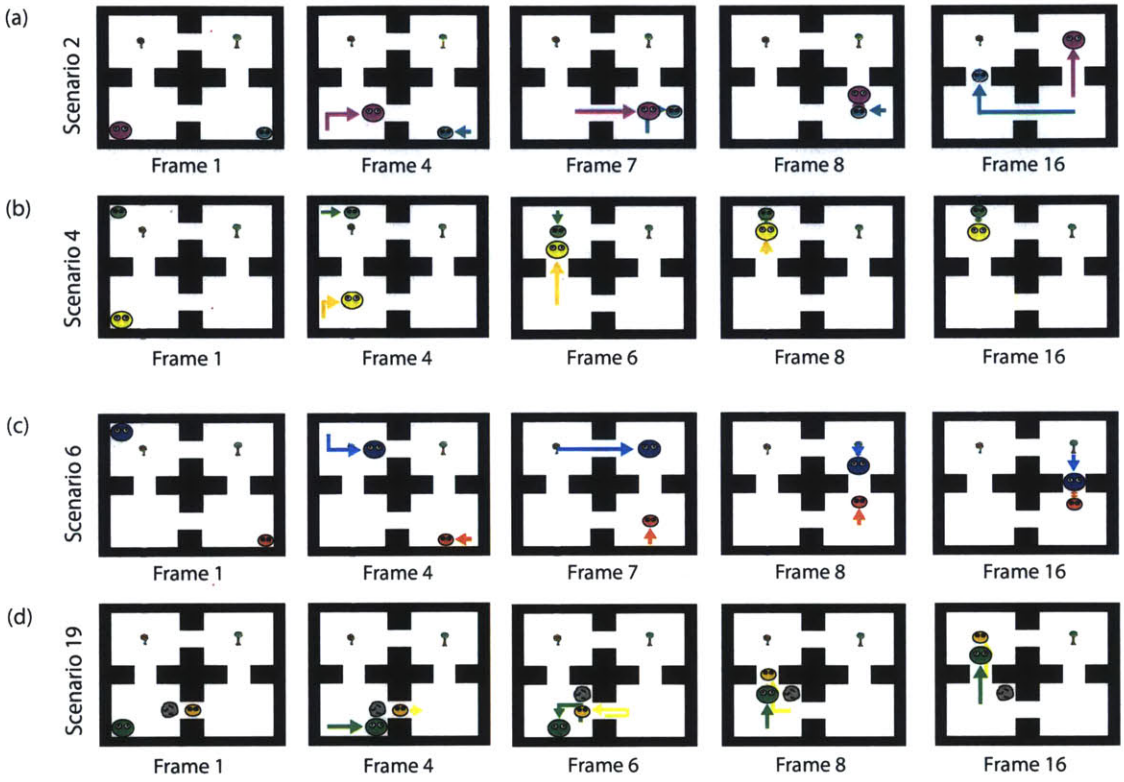


Figure 2-4: Example interactions between Small and Large agents. Agents start as in Frame 1 and progress along the corresponding colored paths. Each frame after Frame 1 corresponds to a *probe point* at which the video was cut off and participants were asked to judge the agents' goals. **(a)** The Large agent moves over each of the goal objects (Frames 1-7) and so the video is initially ambiguous between his having an object goal and a social goal. Disambiguation occurs from Frame 8, when the Large agent moves down and blocks the Small agent from continuing his path up to the object goal. **(b)** The Large agent moves the boulder, unblocking the Small agent's shortest path to the flower (Frames 1-6). Once the Small agent moves into the same room (6), the Large agent pushes him up to flower and allows him to rest there (8-16).

the appropriate space. If the action failed, the agent remained where it was. This allowed participants to perceive the failed actions of the "Small" agent, providing them with information of its possible goal even if it did not succeed in moving. This

corresponds to the experiment described in [75], in which infants could see an agent attempting to move uphill and failing.

We produced videos of 16 frames in length, displaying each scenario. We showed three snippets from each video, which stopped some number of frames before the end. For example, the three snippets of scenario 6 were cut off at frames 4, 7, and 8 respectively (see Fig. 2-4(a)). Participants were asked to make goal attributions at the end of both the snippets and the full 16-frame videos. Asking participants for goal attributions at multiple points in a sequence allowed us to track the change in their judgments as evidence for particular goals accumulated. These cut-off or *probe* points were selected to try to capture key events in the scenarios and so occurred before and after crucial actions that disambiguated between different goals. Since each scenario was used to create 4 stimuli of varying length, we had a total of 96 stimuli.

2.3.3 Procedure

Participants were initially shown a set of familiarization videos of agents interacting in the maze, illustrating the structural properties of the maze-world (*e.g.* the actions available to agents and the possibility of moving obstacles) and the differences between Small and Large agents. The experimental stimuli were then presented in four blocks, each containing 24 videos. Scenarios were randomized within blocks across participants. The left-right orientation of agents and goals was counterbalanced across participants. Participants were told that each snippet would contain two new agents (one Small and one Large) and this was highlighted in the stimuli by randomly varying the color of the agents for each snippet. Participants were told that agents had complete knowledge of the physical structure of the maze, including

the position of all goals, agents and obstacles. After each snippet, participants made a forced-choice for the goal of each agent. For the Large agent, they could select either of the two social goals and either of the two object goals. For the Small agent, they could choose only from the object goals. Participants also rated their confidence on a 3-point scale.

2.3.4 Modeling

Inverse Planning Model

Inverse Planning model predictions were generated using Eq. 2.5, assuming uniform priors on goals, and were compared directly to participants' judgments. In our experiments, the world is given by a 2D maze-world, and the state space includes the set of positions that agents and objects can jointly occupy without overlapping. The set of actions includes *Up*, *Down*, *Left*, *Right* and *Stay* and we assume that $c(A \in \{Up, Down, Left, Right\}) = 1$, and $c(Stay) = 0.1$ to reflect the greater cost of moving than staying put. We set β to 2 and γ to 0.99, following[8].

For the other parameters (namely ρ_g, δ_g and ρ_o) we integrated over a range of values that provided a good statistical fit to our stimuli. For instance, some stimuli were suggestive of "field" goals rather than point goals, and marginalizing over δ_g allowed our models to capture this. Values for ρ_g ranged from 0.5 to 2.5, going from a weak to a strong reward. For δ_g we integrated over three possible values: 0.5, 2.5 and 10.5. These corresponded to "point" object goals (agent receives reward for being on the goal only), "room" object goals (agent receives the most reward for being on the goal and some reward for being in the same room as the goal) and "full space" object goals (agent receives reward at any point in proportion to distance from goal). Values for ρ_o ranged from 1 to 9, ranging from caring weakly about the other agent

to caring about it to a high degree.

Visual Cue Model

We compared the Inverse Planning model to a model that made inferences about goals based on simple visual cues, inspired by previous heuristic- or perceptually-based accounts of human action understanding of similar 2D animated displays [16, 198]. Our aim was to test whether accurate goal inferences could be made simply by recognizing perceptual cues that correlate with goals, rather than by inverting a rational model. We constructed our "Cue-based" model by selecting ten visual cues (listed below), including nearly all the applicable cues from the existing cue-based model described in [16], leaving out those that do not apply to our stimuli, such as heading, angle and acceleration. We then formulated an inference model based on these cues by using multinomial logistic regression to participants' average judgments. The set of cues was as following: (1) the distance moved on the last timestep, (2) the change in movement distance between successive timesteps, (3+4) the geodesic distance to goals 1 and 2, (5+6) the change in distance to goals 1 and 2 (7) the distance to Small, (8) the change in distance to Small, (9+10) the distance of Small to goals 1 and 2.

2.3.5 Results

Our main question is in the psychology of high-level social goals, therefore we analyzed only participants' judgments about the Large agents, which are the ones capable of social goals and complex representations of other agents. Each participant judged a total of 96 stimuli, corresponding to 4 time points along each of 24 scenarios. For each of these 96 stimuli, we computed an empirical probability distri-

bution representing how likely a participant was to believe that the Large agent had each of the four goals 'flower', 'tree', 'help', or 'hinder', by averaging judgments for that stimulus across participants, weighted by participants' confidence ratings. All analyses then compared these average human judgments to the predictions of the Inverse Planning and Cue-based models.

Our main finding is that people's judgments for social and non-social goals matched the Inverse Planning model to a high degree, whereas the Cue-based model was generally unable to distinguish between social goals. The Cue-based model is able to match non-social goal inferences to a high-degree, but at the end of social-goal scenarios it is essentially at chance guessing whether the goal was helping or hindering. This suggests that simple cues such as minimizing distance might be able to guide people's goal inferences when dealing with simple goals, but more abstract reasoning is required for high-level social goals.

Another key finding is that the Inverse-Planning model does equally well on scenarios involving 'boulder' obstacles and scenarios not involving obstacles, whereas the performance of the Cue-based model drops drastically if trained on one set of scenarios and used on another. This shows that cues that were useful in some scenarios for diagnosing 'helping' might become useless in qualitatively scenarios, whereas the basic abstract principle driving the Inverse Planning inference remains equally useful across many different scenarios. This finding echoes the philosophers' point about there being no 'intrinsically moral action' in and of itself. There is no one cue or action feature which can diagnose it as 'helping' for every given scenario - sometimes pushing a boulder towards another agent is helpful, sometimes it is not, depending on the goals of the other agent and the overall environment.

In terms of the linear correlation between human judgment and predictions of the models, the results are as follows: Overall, considering all goal types and training

the Cue-based model on both obstacles and non-obstacles , the two models appear to perform similarly ($r = 0.83$, for the Inverse Planning model, and $r = 0.7$ for the Cue-based model). However, by breaking these correlations down by goal type we find significant differences between the models on social versus object goals (see Fig. 2-5).

The Inverse Planning model correlates well with judgments for all goal types: $r = 0.79, 0.77, 0.86, 0.81$ for flower, tree, helping, and hindering respectively. The Cue-based model correlates well with judgments for object goals ($r = 0.85, 0.90$ for flower, tree) – indeed slightly better the Inverse Planning model – but much less well for social goals ($r = 0.67, 0.66$ for helping, hindering). The most notable differences come on the left-hand sides of the bottom panels in Fig. 2-5. There are many stimuli for which people are very confident that the Large agent is either helping or hindering, and the Inverse Planning model is similarly confident (bar heights near 1). The Cue-based model, in contrast, is unsure: it assigns roughly equal probabilities of helping or hindering to these cases (bar heights near 0.5). In other words, the Cue-based model is effective at inferring simple object goals of maze-world agents, but is generally unable to distinguish between the more complex goals of helping and hindering. When constrained to simply differentiating between social and object goals both models succeed equally ($r = 0.84$), where in the Cue-based model this is probably because moving away from the object goals serves as a good cue to separate these categories. However, the inverse planning model is more successful in differentiating the right goal within social goals ($r = 0.73$ for the inverse planning model vs. $r = 0.44$ for the Cue-based model). Even the slight superiority of the Cue-based model at judging object goals is probably driven by the single cue of getting closer to the target goal, which was particularly useful when the Large agent had an object goal. In these cases the agent always moved directly to it along the shortest

path. It made no errors and never had to take an indirect route. The homogeneity of these cases is favorable to a model based on visual cues. More varied stimuli would make even object-goal judgments more taxing for a visual percept based model (see for example [10]).

Several other general trends in the results are worth noting. The Inverse Planning model fits very closely with the judgments participants make after the full 16-frame videos. On 23 of the 24 scenarios, humans and the inverse planning model have the highest posterior / rating in the same goal ($r = 0.97$, contrasted with $r = 0.77$ for the Cue-based model). It should be noted that in the one scenario for which humans and the inverse planning model disagreed after observing the full sequence, both humans and the model were close to being ambivalent whether the Large agent was hindering or interested in the flower. There is also evidence that the reasonably good overall correlation for the Cue-based model is partially due to overfitting; this should not be surprising given how many free parameters the model has. We divided scenarios into two groups depending on whether a boulder was moved around in the scenario, since movable boulders increase the range of variability in helping and hindering action sequences. When trained on the 'no boulder' cases, the Cue-based model correlates poorly with participants average judgments on the 'boulder' cases: $r = 0.42$. The same failure of transfer occurs when the Cue-based model is trained on the 'boulder' cases and testing on the 'no boulder' cases: $r = 0.36$ on the test stimuli. As discussed above, this is consistent with our general concern that a Cue-based model incorporating many free parameters may do well when tailored to a particular environment, but is not likely to generalize well to new environments. In contrast, the Inverse Planning model captures abstract relations between the agents and their possible goal and so lends itself to a variety of environments without requiring a growing number of parameters.

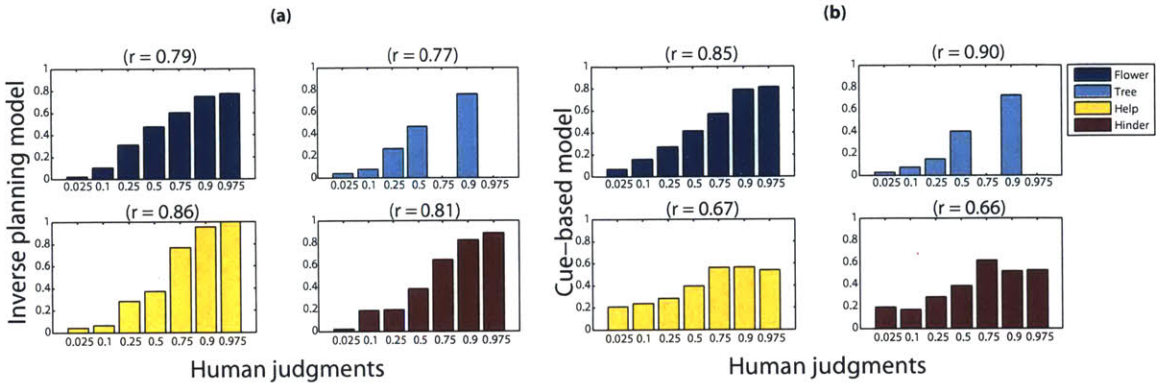


Figure 2-5: Correlations between human goal judgments and predictions of the Inverse Planning model **(a)** and the Cue-based model **(b)**, broken down by goal type. Bars correspond to bins of stimuli (out of 96 total) on which the average human judgment for the probability of that goal was within a particular range; the midpoint of each bin's range is shown on the x-axis labels. The height of each bar shows the model's average probability judgment for all stimuli in that bin. Linear correlations between the model's goal probabilities and average human judgments for all 96 stimuli are given in the y-axis labels.

The inability of the heuristic model to distinguish between helping and hindering is illustrated by the plots in Fig. 2-6. In contrast, both the Inverse Planning model and the human participants are often very confident that an agent is helping and not hindering (or vice versa).

Fig. 2-6 also illustrates a more general finding, that the Inverse Planning model captures most of the major qualitative shifts (e.g. shifts resulting from disambiguating sequences) in participants' goal attribution. Figure 2-6 displays mean human judgments on four scenarios. Probe points (i.e. points within the sequences at which participants made judgments) are indicated on the plots and human data is compared with predictions from the Inverse Planning model and the Cue-based model.

On scenario 6 (depicted in Fig. 2-4(a) but with goals switched), both the Inverse Planning model and humans participants recognize the movement of the Large agent

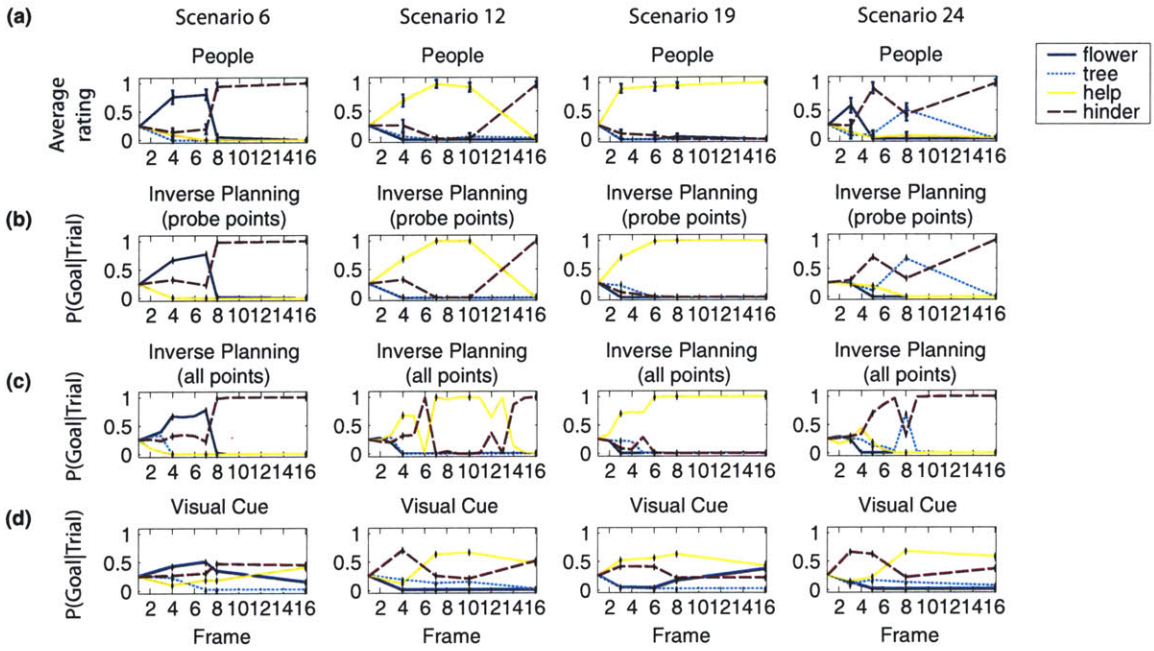


Figure 2-6: Example data and model predictions. Probe points are marked as black circles. **(a)** Average participant ratings with standard error bars. **(b)** Predictions of Inverse Planning model interpolated from cut points. **(c)** Predictions of Inverse Planning model for all points in the sequence. **(d)** Predictions of Cue-based model.

one step off the flower (or the tree in Fig. 2-4(b)) as strong evidence that Large has a hindering goal. The Cue-based model responds in the same way but with much less confidence in hindering. Even after 8 subsequent frames of action it is unable to decide in favor of hindering over helping.

While the Inverse Planning model and participants almost always agree by the end of a sequence, they sometimes disagree at early probe points. In scenario 5, both agents start off in the bottom-left room, but with the Small agent right at the entrance to the top-left room. As the Small agent tries to move towards the flower (the top-left goal), the Large agent moves up from below and pushes Small one step towards the flower before moving off to the right to the tree. People interpret the

Large agent's action as strong evidence for helping, in contrast with the Inverse Planning model. For the model, because Small is so close to his goal, Large could just as well stay put and save his own action costs. Therefore his movement upwards is not evidence of helping.

2.4 General Discussion

There is nothing either good or bad, but thinking makes it so

– Hamlet, Act II, Scene II

Our goal in this chapter was to address two challenges. The first challenge was to formalize social goal attribution within a general theory-based model of intuitive psychology. This model had to account for the general range of behaviors that humans judge as evidence of helping or hindering. The second, more specific challenge was for the model to perform well on a demanding inference task in which social goals must be inferred from very few observations without direct evidence of the agents' goals.

The experimental results go some way toward meeting these challenges. The Inverse Planning model classified a diverse range of agent interactions as helping or hindering in line with human judgments. This model also distinguished itself against a model based solely on simple perceptual cues. It produced a closer fit to humans for both social and nonsocial goal attributions, and was far superior to the visual cue model in discriminating between helping and hindering.

The essential extension to previous work on action understanding as inverse-planning is the addition of a *Principle of Sympathy.* Much like the Principle of Rational Action, this notion can abstract away many details about any specific sce-

nario and provide general guidelines for goal achievement. By assuming other agents behave according to this principle, a rational observer can understand a myriad of potentially novel situations. The principle is captured in computational terms by recursive utility functions in multi-agents MDPs. A "helpful" agent adopts the utility of others as its own, and a "hindering" agent adopts the negative of this utility. This echoes the etymology of the word 'sympathy' itself, made up of the words 'feeling' and 'together'.

While our experiments were conducted with adults, our model is well-suited to capture the findings come infant literature, such as [75, 76]. In almost all of these infant experiments the actions of the helpers and hinderers were perceptually distinct: for example, hindering agents pushed downhill and closed boxes, helpful agents pushed uphill and opened boxes. This leaves open the possibility that infants are using cue-based perceptual models to classify social agents. New joint work [93] shows that infants distinguish perceptually identical actions depending on the social goals, preferences, and perceptual access of other agents, as predicted by an extension of the model presented here.

This new work required the model to go beyond a simplifying assumption made here - that the knowledge of the observer is shared by the social agents. But once false belief or different states of knowledge are possible, social and moral evaluations become more markedly more complex. Suddenly one can have scenarios like "Alice thought Bob was bad and tried to hinder him, however Bob was good, but Alice did not know this and so does not deserve punishment for hindering", or "Alex is Beth's friend and wants the best for her. He knows Beth wants The Thing, but thinks that this is foolish and that if only Beth really knew what was good for her, she would not want The Thing. Alex did not help Beth get The Thing". More complex, certainly, but also more realistic.

Uncertainties over the true nature of things apply to goals and intentions, but also to the physical world itself. Having such uncertainties at all levels of the model means that inferences can be made at different levels too. For example, if you are unsure about an agent's intention and observe a highly diagnostic action (like smacking), you may draw strong conclusions about the agent's intentions (to harm). But a high degree of certainty about the intention could instead drive conclusions about elements of the world. Think back to the case of the child reaching for a hot stove and receiving a smack on the hand from her mother. If the child has a high certainty that the mother is trying to help her, she would infer new knowledge about the world that caused her mother's actions: the stove is dangerous. Much of pedagogy is supported by the assumption that the teacher is not only knowledgeable, but also trying to be helpful in explaining the world [46].

These complexities are a challenge for people, not just models. In general, the 'inverse' direction of Bayesian inference is hard. In vision, for example, there is a large space of possible 'scenes' that can produce the same visual percept. In social contexts, there is a large space of social and moral properties that can explain a sequence of events. In visual perception there is generally an agreed upon 'solution' in the form of a high-probability visual scene all people converge on [2]. But in moral and social inferences – and perhaps in all high-level cognition – there is no agreed upon 'single solution'. There may be several competing and incompatible 'high probability solutions', some appealing to intentions, others to beliefs, others to the world.

The cue-based approach often cites the automatic nature of certain moral and social evaluations as evidence that the inference process is similar to a bottom-up perceptual problem. We agree that visual perception and inference of intentions and

[2]Barring certain illusions, and even then there are only a few common percepts

goals share certain basic mechanisms, however the results presented here suggest that what they might share is the underlying mechanism of Bayesian inference. - Vision has been highly optimized over evolutionary time, but social perception was not, perhaps cannot be. We do not doubt the reality of some cues for animacy and simple social evaluation, but these might serve only to bias people as part of their otherwise more mentalistic evaluation.

Whoever draws too close, and catches the Sirens' voice in the air: no sailing home for him, no wife rising to meet him, no happy children beaming up at their father's face (The Odyssey)

The moral and social explanations people use can be contested and are subject to revision and change upon reflection, a hallmark of high-level processes. The naive person suddenly appearing on Odysseus' ship may have thought the sailors were treating Odysseus cruelly based on what she saw with her own eyes. But having read through the *Odyssey*, she might change her mind.

Chapter 3

Learning Physics*

There are children playing in the street who could solve some of my top problems in physics, because they have modes of sensory perception that I lost long ago. — J. Robert Oppenheimer, as quoted by Marshall Mcluhan

3.1 Introduction

Reasoning about the physical properties of the world around us is a ubiquitous feature of human mental life. Not a moment passes when we are not, at least at some implicit level, making physical inferences and predictions. Glancing at a book on a table, we can rapidly tell if it is about to fall, or how it will slide if pushed, tumble

*Joint work with Andreas Stuhlmller, Noah Goodman and Josh Tenenbaum

if it falls on a hard floor, sag if pressured, bend if bent. The capacity for physical scene understanding begins to develop early in infancy, and has been suggested as a core component of human cognitive architecture [168].

While some parts of this capacity are likely innate [7], learning also occurs at multiple levels from infancy into adulthood. Infants develop notions of containment, support, stability, and gravitational force over the first few months of life [124, 3], as well as differentiating between liquid substances and solid objects [80]. Young children build an intuitive understanding of remote controls, touch screens, magnets and other physical devices that did not exist over most of our evolutionary history. Astronauts and undersea divers learn to adapt to weightless or partially weightless environments [118], and videogame players can adjust to a wide range of game worlds with physical laws differing in some way from our everyday natural experience.

Not only can we learn or adapt our intuitive physics, but we can often do so from remarkably limited and impoverished data. While extensive experience may be necessary to achieve expertise and fluency, only a few exposures are sufficient to grasp the basics of how a touch screen device works, or to recognize the main ways in which a zero-gravity environment differs from a terrestrial one. While active intervention and experimentation can be valuable in discovering hidden causal structure, they are often not necessary; observation alone is sufficient to infer how these and many aspects of physics operate. People can also can gain an intuitive appreciation of physical phenomena which they can only observe or interact with indirectly, such as the dynamics of weather fronts, ocean waves, volcanoes or geysers.

Several questions naturally follow. How, in principle, can people learn aspects of intuitive physics from experience? What is the form of the knowledge that they learn? How can they grasp structure at multiple levels, ranging from deep enduring laws acquired early in infancy to the wide spectrum of novel and unfamiliar dynamics

that adults encounter and can adapt to? How much and what kind of data are required for learning different aspects of physics, and how are the data brought to bear on candidate hypotheses? In this chapter we present a theoretical framework that aims to answer these questions in computational terms, and a large-scale behavioral experiment that tests the framework as an account of how people learn basic aspects of physical dynamics from brief moving scenes.

Our modeling framework takes as a starting point the computational-level view of theory learning as rational statistical inference over hierarchies of structured representations [178, 68]. Previous work in this tradition focused on relatively spare and static logical descriptions of theories and data; for example, a law of magnetism might be represented as 'if magnet(x) and magnet(y) then attract(x,y)', and the learner's data might consist of propositions such as 'attracts($object_a$, $object_b$)' [92]. Here we adopt a more expressive representational framework suitable for learning the force laws and latent properties governing how objects move and interact with each other, given observations of scenes unfolding dynamically over time. Our representation includes both logical machinery to express abstract properties and laws, but also numerical and vector resources needed to express the observable trajectories of objects in motion, and the underlying force dynamics causally responsible for those motions. We can express all of this knowledge in terms of a *probabilistic program* in a language such as Church [58, 59].

An example of the kind of dynamic scenes we study is shown in Fig. 3-1. Imagine this as something like an air hockey table viewed from above. There are four disk-shaped "pucks" moving in a two-dimensional rectangular environment under the influence of various causal laws and causally relevant properties. In a physical domain the causal laws are *force* laws, and these forces may be either local and pairwise (analogous to the way two magnetic objects typically interact) or global (analogous

to the way gravity operates in our typical environment). The properties are physical properties that determine how forces act on objects, and may include both object-based and surface-based features, analogous to inertial mass and friction respectively. A child or adult looking at such a display might come to a conclusion such as 'red pucks attract one another' or 'green patches slow down objects'. With the right configuration different physical properties begin to interact, such that an object might be seen as heavy, but in the presence of a patch that slowed it down its 'heaviness' might be explained away as friction.

Such dynamical displays are still far simpler than the natural scenes people see early in development, but they are much richer than the stimuli that has been studied in previous experiments on learning intuitive physics and learning in intuitive causal theories more generally. Previous research on learning physics from dynamical scenes has tended to focus on the inference of object properties under known force laws, and typically on only the simplest case: inferring a single property from a single dynamical interaction, as in inferring the relative mass of two objects from observing a single collision between them with one object starting at rest (see for example [140, 55, 179, 2]).

Some research on causal learning more generally has looked at the joint inference of causal laws and object attributes, but only in the presence of simple discrete events rather than a rich dynamical scene [65, 67, 71]. For example, from observing that a "blicket-detector" lights up when objects A or B are placed on it alone or together, but does not light up when objects C or D are placed on it alone or in combination with A or B, people may infer that only objects A and B are blickets, and that the blicket detector only lights up when all the objects on it are blickets [109]. It is not clear that studying how people learn from a small number of isolated discrete events presented deliberately and pedagogically generalizes to how they learn physics in the

real world, where configurations of objects move continuously in space and time and interact in complex ways that are hard to demarcate or discretize.

In this sense our experiments are intended to capture much more of how we learn physics in the real world. Participants observe multiple objects in motion over a period of five seconds, during which the objects typically collide multiple times with each other as well as with stationary obstacles, pass over surfaces with different frictional properties, and move with widely varying velocities and accelerations. We compare the performance of human learners in these scenarios with the performance of an ideal Bayesian learner who can represent precisely the dynamical laws and properties at work in these stimuli. While people are generally able to perform this challenging task in ways broadly consistent with an ideal observer model, they also make systematic errors which are suggestive of how they might use feature-based inference schemes to approximate ideal Bayesian inference. Hence we also compare people's performance to a hybrid model that combines the two kinds of inference (ideal and feature-based), suggesting how to build a unified account which is based both on heuristics and an implicit understanding of Newtonian-like mechanics.

3.2 Formalizing Physics Learning

The core of our formal treatment is a hierarchical probabilistic generative model for theories [92, 187, 62], specialized to the domain of intuitive physical theories (Fig. 3-2). The hierarchy consists of several levels, with more concrete (lower-level) concepts being generated from more abstract versions in the level above, and ultimately bottoming out in data that take the form of dynamic motion stimuli.

Generative knowledge at each level is represented formally using `(define ...)` statements in Church, a stochastic programming language [58]. The `(define x v)`

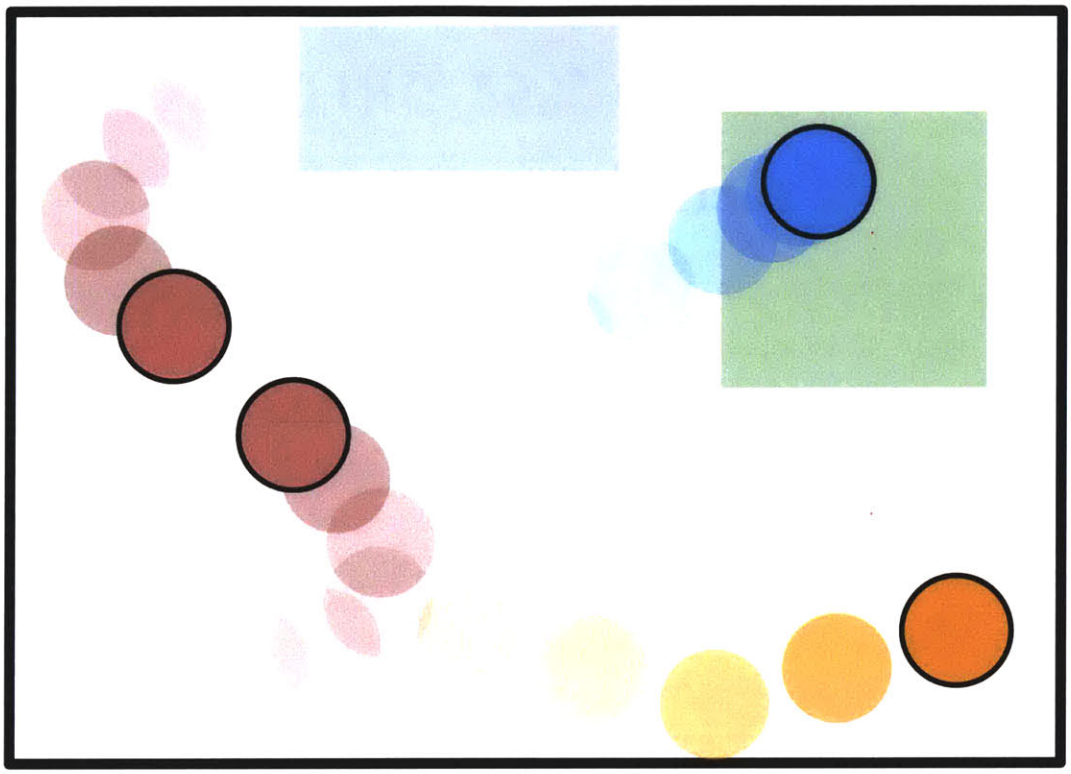


Figure 3-1: Illustration of the domain explored in this chapter, showing the motion and interaction of different pucks moving on a two-dimensional plane governed by latent physical properties and dynamical laws, such as mass, friction, global forces and pairwise forces.

statement binds the value v to the variable x, much as the statement `a = 3` binds the value 3 to the variable a in many programming languages. In probabilistic programming, however, we often bind variables with values that come from probability distributions, and thus on each run of the program the variable might have a different value. For example, `(define dice (uniform-draw 1 6))` stochastically assigns a value between 1 and 6 to the variable `dice`. Whenever the program is run, a different value is sampled and assigned to `dice`, drawing from the uniform distribution.

Probabilistic programs are useful for representing knowledge with uncertainty

(see for example [58, 172, 60]). Fig. 3-2(iii) shows examples of probabilistic definition statements within our domain of intuitive physics, using Church. Fig. 3-2(i) shows the levels associated with these statements, and the arrows from one level to the next show that each level is sampled from the definitions and associated probability distributions of the level above it. The definition statements provide a formalization of the main parts of the model. The full forward generative model is available at `http://forestdb.org/models/learning-physics.html`

In the text below we will explain these ideas further, using informal English descriptions whenever possible, but see [58] for a more formal treatment of the programming language Church, and probabilistic programming in general.

Framework level. The top-most level N represents general framework knowledge [191] and expectations about physical domains. The concepts in this level include **entities**, which are a collection of **properties**, and **forces**, which are functions of properties and govern how these properties change over time. Forces can be fields that apply uniformly in space and time, such as gravity, or can be event-based, such as the force impulses exerted between two objects during a collision or the forces of kinetic friction between two objects moving over each other.

All nature, then, as self-sustained, consists / Of twain of things: of bodies and of void / In which they're set, and where they're moved around.
– Lucretius, On the Nature of Things

Properties are named values or distributions over values. While different entities can have any number of properties, a small set of properties are 'privileged': it is assumed all entities have them. In our setup, the properties *location* and *shape* are privileged in this sense.

Entities are further divided into 'static' and 'dynamic'. Dynamic entities are those that can potentially move, and all dynamic entities have the privileged property *mass*. Dynamic entities correspond then to the common sense definition of matter as 'a thing with mass that occupies space' [1].

[1]The static/dynamic distinction is motivated by similar atomic choices in most computer physics

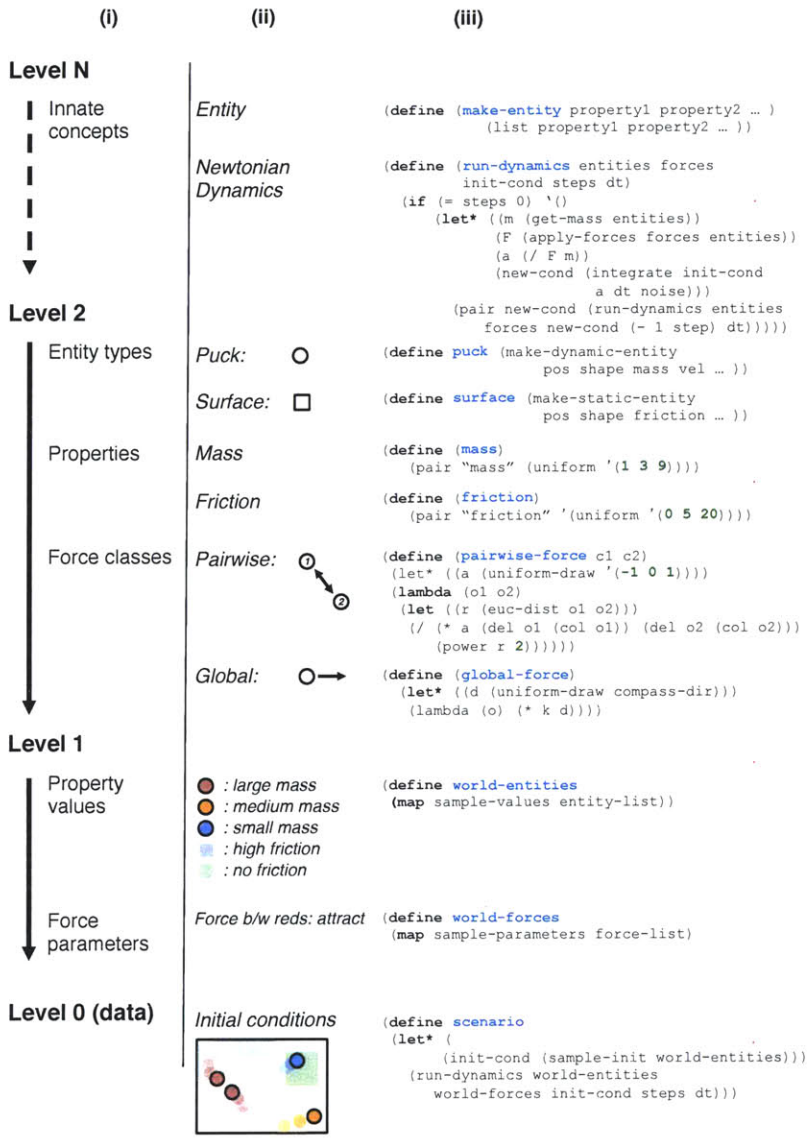


Figure 3-2: Formal framework for learning intuitive physics in different domains: (i) The general hierarchy going from abstract principles and assumptions to observable data. The top-most level of the hierarchy assumes a general noisy-Newtonian dynamics. (ii) Applying the principles in the left-most column to the particular domain illustrated by Fig. 3-1 (iii) Definition statements in Church, capturing the notions shown in the middle column with a probabilistic programming language.

engines used for approximate dynamic simulations, engines that were suggested as models of human intuitive physics (e.g. [12]). In these physics engines the static/dynamic divide allows computational speed-up and memory conservation, since many forces and properties don't have to be calculated or updated for static entities. It is an interesting possibility that the same kind of short-cuts developed by engineers trying to quickly simulate physical models might also represent a cognitive distinction. Similar notions have been proposed in cognitive development in the separation of 'objects' from more stable 'landscapes' [106]

The framework level defines a 'Newtonian-like' dynamics, where acceleration is proportional to the sum of the forces acting on an object's position relative to the object's mass. This is consistent with suggestions from several recent studies of intuitive physical reasoning in adults [12, 164, 54, 144] and infants [174]. As [144] show, such a 'noisy-Newtonian' representation of intuitive physics can account for previous findings in dynamical perception that have supported a heuristic account of physical reasoning [55, 56, 179], or direct perception models [140, 2].

Descending the hierarchy. Descending from Level N to Level 0, concepts are increasingly grounded by sampling from the concepts and associated probability distributions of the level above (Fig. 3-2(i)). Each level in the hierarchy can spawn a large number of instantiations in the level below it. Each lower level of the hierarchy contains more specific entities, properties and forces than the level above it. An example of moving from Level N to Level N-1 would be grounding the general concepts of entities and forces as more specifically 2-dimensional masses acting under collisions. An alternative would ground the same general entities and forces as 3-dimensional masses acting under conservation forces. This grounding can proceed through an indeterminate number of levels, until it ultimately grounds out in observable data (Level 0).

Space of learnable theories. Levels 0-2 in Fig. 3-2 capture the specific subdomain of intuitive physics we study in this chapter's experiments: two-dimensional discs moving over various surfaces, generating and being affected by various forces, colliding elastically with each other and with barriers bounding the environment (cf Fig. 3-1).

Levels 0-2 represent the minimal framework needed to explain behavior in our task and we remain agnostic about more abstract background knowledge that might also be brought to bear. We give participants explicit instructions that help determine

a single Level 2 schema for the task, which generates a large hypothesis space of candidate Level 1 theories, which they are asked to infer by using observed data at Level 0.

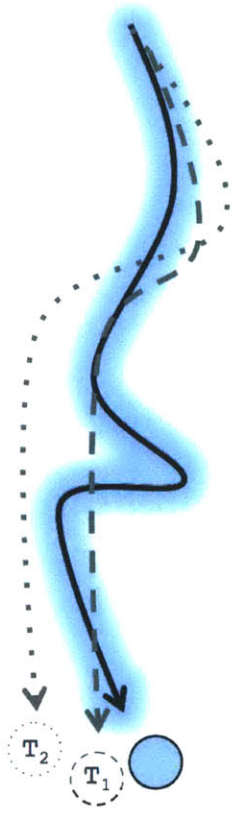
Level 2: The "hockey-puck" domain. This level specifies the entity types *puck* and *surface*. All entities within the type *puck* have the properties *mass*, *elasticity*, *color*, *shape*, *position*, and *velocity*. Level 2 also specifies two types of force: *Pairwise forces* cause pucks to attract or repel, following the 'inverse square distance' form of Newton's gravitation law and Coulomb's Law. *Global forces* push all pucks in a single compass direction. We assume forces of *collision* and *friction* that follow their standard forms, but they are not the subject of inference here.

Level 1: Specific theories. The hockey-puck domain can be instantiated as many different specific theories, each describing the dynamics of a different possible world in this domain. A Level 1 theory is determined by sampling particular values for all free parameters in the force types, and for all entity subtypes and their subtype properties (e.g., masses of pucks, friction coefficients of surfaces). Each of the sampled values is drawn from a probability distribution that the Level 2 theory specifies. So, Level 2 generates a prior distribution over candidate theories for possible worlds in its domain.

The domain we study here allows three types of pucks, indexed by the colors red, blue and yellow. It allows three types of surfaces (other than the default blank surface), indexed by the colors brown, green and purple. Puck mass values are 1, 3, or 9, drawn with equal probability. Surface friction coefficients values are 0, 5 or 20, drawn with equal probability. Different pairwise forces (attraction, repulsion, or no interaction) can act between each of the different pairs of puck types, drawn with equal prior probability. Finally, a global force may push all pucks in a given direction, either $\uparrow, \downarrow, \leftarrow, \rightarrow$ or 0, drawn with equal probability. We further restrict

this space by considering only Level 1 theories in which all subclasses differ in their latent properties (e.g. blue, red and yellow pucks must all have different masses). While this restriction (together with the discretization) limits the otherwise-infinite space of theories, it is still a very large space, containing 131,220 distinct theories [2].

Level 0: Observed data. The bottom level of our hierarchical model (Fig. 3-2) is a concrete scenario, specified by the precise individual entities under observation and the initial conditions of their dynamically updated properties. Each Level 1 theory can be instantiated in many different scenarios. The pucks' initial conditions were drawn from a zero-mean Gaussian distribution for positions and a Gamma distribution for velocities, and filtered for cases in which the pucks began in overlap. Once the entities and initial conditions are set, the positions and velocities of all entities are updated according to the Level 1 theory's specific force dynamics for T time-steps, generating a path of multi-valued data points, $d_0, \ldots, d_T$. The probability of a path is simply the product of the probabilities of all the choices used to generate the scenario. Finally, the actual observed positions and velocities of all entities are assumed to be displaced from their true values by Gaussian noise.



Theory 1 generates a path closer to the (noisy) data and will have a higher posterior than Theory 2

3.2.1 Learning Physics as Bayesian inference

Having specified our overall generative model, and the particular version of it underlying our "hockey puck" domain, we now turn to the question of learning. The model described so far allows us to formalize different kinds of learning as inference over different levels of the hierarchy. This approach can in principle be used for reasoning about all levels of the hierarchy, including the general shape of forces

[2] More precisely, the cross product $N(mass)! \times N(frictioncoefficients)! \times N(direction) \times N(pairwisecombination)^{N(forceconstant)} = 131,220$. Selecting the right theory in this space is equivalent to correctly choosing 17 independent binary choices

and types of entities, the unobserved physical properties of entities, as well as the existence, shape and parameters of unseen dynamical rules. Given observations, an ideal learner can invert the generative framework to obtain the posterior over all possible theories that could have produced the observed data. We then marginalize out nuisance parameters (other irrelevant aspects of the theory) to obtain posterior probabilities over the dynamic quantity of of interest.

Inference at multiple levels includes both continuous parameter estimation (e.g. the strength of an inverse-square attractive force or the exact mass value of an object) and more discrete notions of structure and form (e.g. the very existence and shape of an attractive force, the fact that an object has a certain property). This parallels a distinction between two modes of learning that appears in AI research as well as cognitive development (where it is referred to as "parameter setting" and conceptual change [25]). In general, inferring structure and form (or conceptual change) is seen as harder than parameter estimation.

Learning at different levels could unfold over different spans of time depending on the size and shape of the learning space, as well as on background knowledge and the available evidence. Estimating the mass of an object from a well-known class in a familiar setting could take adults under a second, while understanding that there is a general gravitational force pulling things downwards given little initial data might take infants several months to grasp [95].

In this chapter we consider learning at a mid-point between these two extremes, between inferring basic physical knowledge and estimating parameters in a familiar environment. Our experiments involve joint estimation of multiple parameters and basic structure learning in the form of discrete structural relations (pairwise and global forces), but not the more abstract conceptual change that could take longer and require more evidence. The basic structure of noisy Newtonian mechanics is

assumed present, and we examine learning at Level 1 - the sort of learning that could happen over several seconds in a novel setting.

3.2.2 Simulation based approximations and summary statistics

The Bayesian inversion of the generative model is in principle sufficient for inference over any unknown quantity of interest in it. However, it can be computationally demanding. In this section we consider a psychologically plausible approximation to the generative model, one which combines summary statistics and the ability to imagine new dynamic scenes.

In our experiments, each scenario contained exactly 4 pucks and 2 surfaces. This restricts the number of hypotheses we need to consider to a maximum of 14,580 for any one scenario, out of the larger in-principle space of 131,220. We can sum over all the hypotheses in this domain, but such an approach is not practical for scaling to larger domains and considering their the full hypothesis space, where integration is generally intractable. Even for our restricted domain it is not psychologically plausible a-priori that for any given dynamic stimuli people carry out massive inference over all possible models that could have generated it, given the short time-frame in which they can make judgments. Further, people can use more than local-path information to assess different physical parameters. For example, if people think two objects attract they might reasonably expect that over time the mean distance between the objects should shrink.

This psychological intuition suggests a principled way of approximating the full inference, following a statistical method known as Approximate Bayesian Computation (see [15] for a review). This approach is similar to 'indirect inference' [70],

which assumes a model that can generate simulated data d' given some parameters θ, but does not try to estimate θ directly from observed data d. Rather, we first construct an auxiliary model with parameters β and an estimator $\tilde{\beta}$ that can be evaluated on both d and d'. The indirect estimate of the parameter of interest, $\hat{\theta}$, is then the parameter that generated the simulated data whose estimator value $\tilde{\beta}(d')$ is as close as possible to the estimator value of observed data, $\tilde{\beta}(d)$ (for additional technical details see for example] [70]).

Here we will use the following approximation: Our framework can generate simulated object paths given physical parameters θ, which we then wish to estimate. We begin by drawing simulated data for all the models within the domain over all scenarios, giving us several hundred thousand paths. For every physical parameter θ we construct a set of summary statistics that can be evaluated on any given path, and act as estimators. For example, the summary statistic $avgPositionX(d)$ calculates the mean x-axis position of all objects over a given path, and can be used as an estimator for the existence of a global force along the x-axis. We evaluate these summary statistics for each of the parameter values over all the paths, obtaining an empirical likelihood distribution which is smoothed with Gaussian kernels. The estimated likelihood of a given parameter is then the likelihood of the summary statistic for the observed data (see Fig. 3-3(a) and (b) for an illustration of this process).

Psychologically, this approximation corresponds to the following: people can imagine dynamical scenes unfolding over time, but when reasoning about a specific scene they do not imagine how the same scene could have unfolded under all the different unknown variables they are reasoning about. Instead, they compute some simple summary statistics of the specific scene, e.g. how close are some pucks on average. People then compare the value of these summary statistics to a repository which was calculated over many possible scenes. These repositories are built

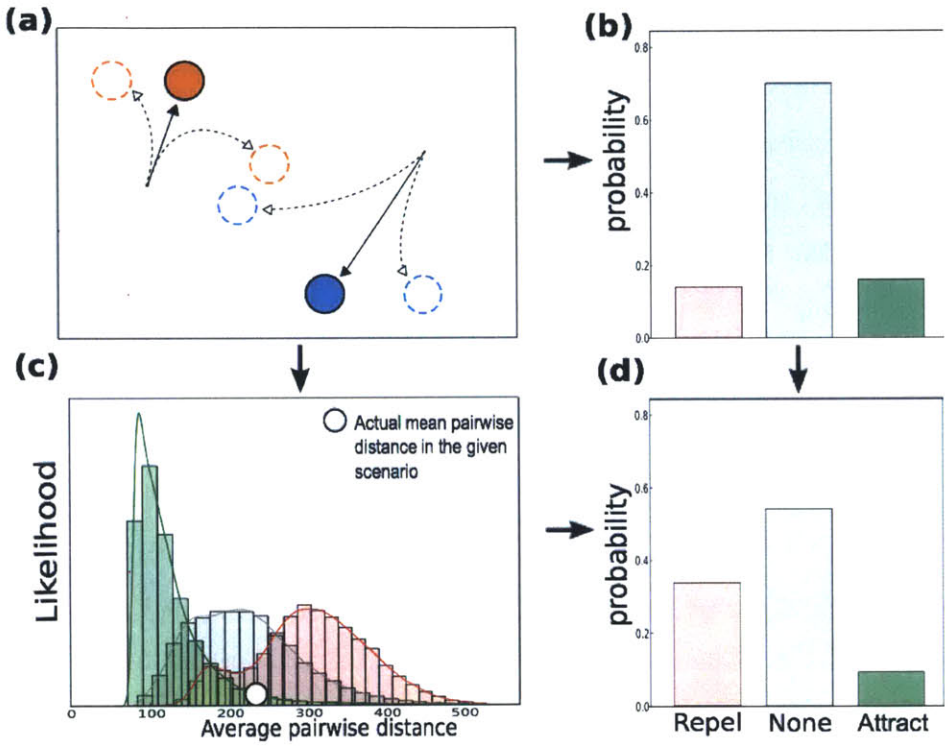


Figure 3-3: Approximations and the ideal observer for pairwise forces. For a given scenario (a), many alternate paths are generated and compared to the observed path, giving us a log likelihood for all theories. Posterior estimates are obtained by either marginalizing over all theories (b), or by comparing the summary statistics of the scenario to its empirical distribution over many simulations (c). We also consider a simple combination of the methods (d).

up by using the same imagery capacities which allow people to imagine individual scenes evolving over time, possibly in an off-line manner (as was the case in our models). This approximation relies on imagery, imagination and simulation, rather than obtaining direct experience of tens of thousands of different scenarios and building different features to use as classifiers of theories.

Our set of summary statistics included average position and total change along the x-axis, average position and total change along the y-axis, mean pairwise distance

between particles, total change in mean pairwise distance, average velocity, velocity loss while on surfaces, amount of time spent at rest while on surfaces and the ratio of pre- and post-collision velocities [3]. These summary statistics are meant to capture a large amount of possible perceptual data in the stimuli, but they are not meant to be exhaustive. We take up the question of possible summary statistics again in the general discussion.

While indirect inference and approximation techniques are useful, they have certain limitations, such as being insensitive to the particular conditions in outlying scenarios. That is, for any given summary statistic it is easy to construct a simple scenario which is unlikely under the statistic's likelihood, and yet people will be able to reason about without difficulty. An interesting possibility is to combine the strengths of the ideal observer model described in the previous section together with summary statistics. Below we will consider for simplicity combinations of the likelihoods derived from each approach.

Finally, we stress that this approximation technique is not an alternative to the idea of inference through simulation, but rather a potentially necessary supplement to it. The simulation-based approach and related approximation is in contrast to a different possible way of approximately scoring theories, which is to learn through experience associations between theories and many features. This would require a great deal of experience indeed, which people are unlikely to come by for the synthetic scenarios considered here for example. This contrast is similar to the debates about top-down vs. bottom-up techniques in object perception, between those who stress a more top-down approach that relies an on actual 3D object model,

[3]The velocity statistic was chosen based on heuristic models suggesting people are sensitive to this data [56, 55]. The change in angle following a collision was also considered based on this work, but it was found to actually be negatively correlated with mass judgments, which is in line with the findings of [140].

and those who stress bottom-up perceptual cues calculated from still images and used for classification.

We now examine these various ways of physical reasoning, by considering people's performance on a novel dynamical task.

3.3 Experiment

3.3.1 Participants

Three hundred participants from the US were recruited via the Amazon Mechanical Turk service, and were paid for their participation. Ten participants were excluded from analysis for failing comprehension questions.

3.3.2 Stimuli

60 videos were used as stimuli, each lasting 5 seconds and depicting the dynamics of several pucks moving and colliding.

We constructed the stimuli in the following manner: First, we defined a set of 10 *worlds* that differ in the physical rules underlying their dynamics, as well as in the properties of the objects that appear in them. For example: in $world_1$ blue pucks have a large mass and there are no global or coupling forces, whereas in $world_5$ blue pucks are light and red pucks repel one another. A full description of the underlying physical rules of each world is available at `http://www.mit.edu/~tomeru/physics2014/underlyingRules.pdf`

Next, for each world we created 6 different *scenarios* that differ in their initial conditions (i.e. the starting location and velocity of the pucks and surfaces), as well as the particular objects used and the size of the surfaces. For example: the

first scenario of $world_1$ has red, yellow and blue pucks, whereas the third scenario uses only red and yellow pucks. The initial conditions were drawn from random distributions, and in practice most of the movies started with the pucks already moving.

Using the dynamical rules of the world and starting from the initial conditions, we unfolded the scenarios over 400 steps and created a video detailing the motion of the objects over time [4]. All stimuli used are available at `http://www.mit.edu/~tomeru/physics2014/stimuli/`, and a static visual representation is shown in Fig. 3-4 and 3-5.

3.3.3 Procedure

Each participant saw 5 videos drawn from the set of 60 possible stimuli. The video-participant pairing was done according to a Latin-square design, such that approximately thirty participants saw each video. The order of the 5 videos was randomized for each participant.

Participants were informed what objects, forces and physical properties were potentially present across all the stimuli, and also that objects of the same color have the same properties. It was explained that objects can be heavy, medium or light, and that each object type can potentially exert forces on other types: object types either attract, repel or don't interact with one another. Participants were instructed to think of the videos as similar to 'hockey pucks moving over a smooth white table-top', and informed that patches on the plane can have different roughness. Finally, they were told there may or may not be a global force in the world, pulling all objects in a particular direction (north, south, east or west). An example experiment

[4] We used the classical Runge-Kutta method (RK4) for numerical integration to move the entities forward in time.

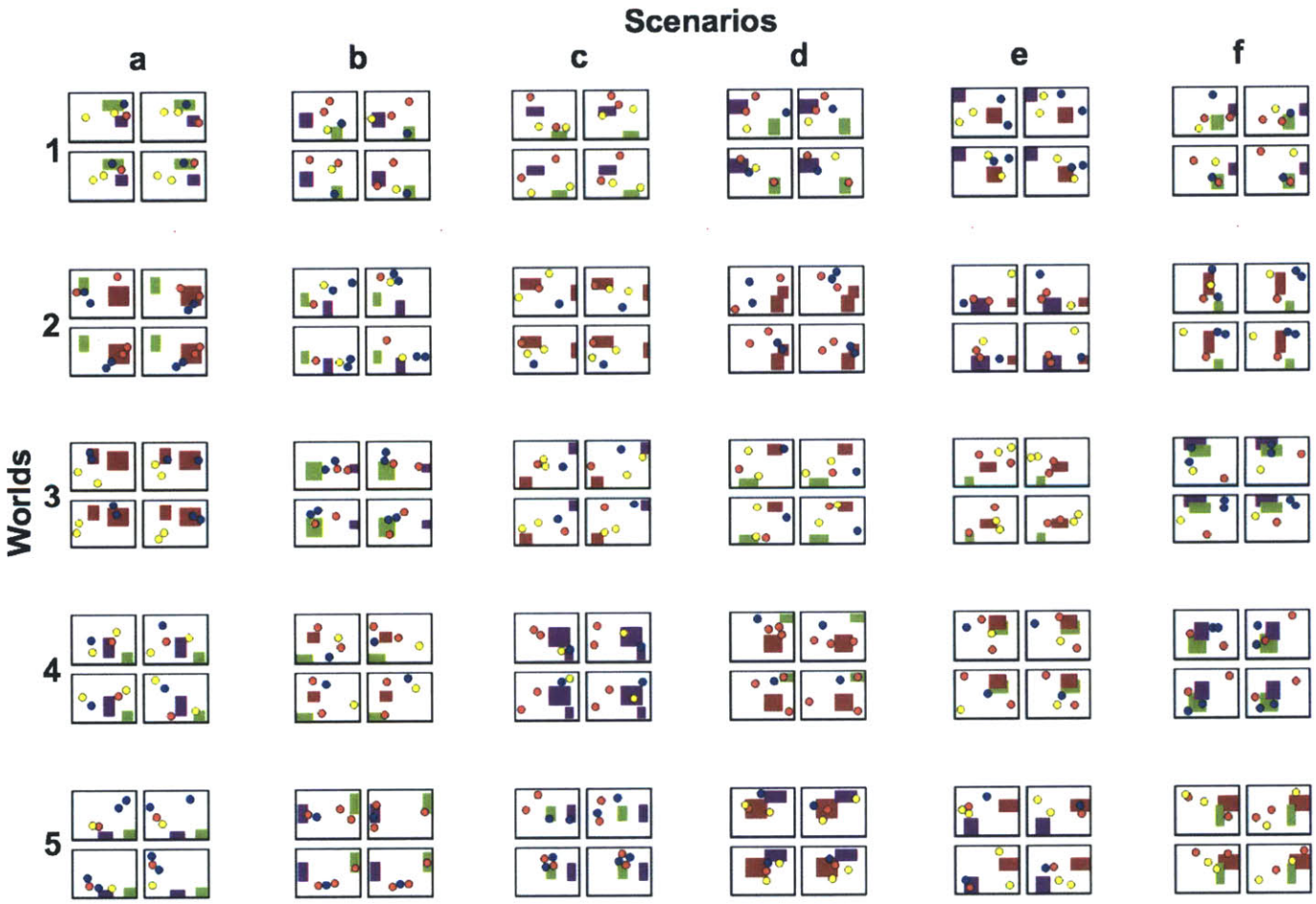


Figure 3-4: Part 1 of all the stimuli used, showing 'worlds' 1-5 with 6 scenarios per world. There are 4 images per scenario, showing it unfold over time. The images were sampled at the start each scenario (upper left image in each scenario), 1.25 seconds into the scenario (upper right image), 3.75 seconds into the scenario (lower left image) and at the end of the scenario (5 seconds after it started, lower right image).

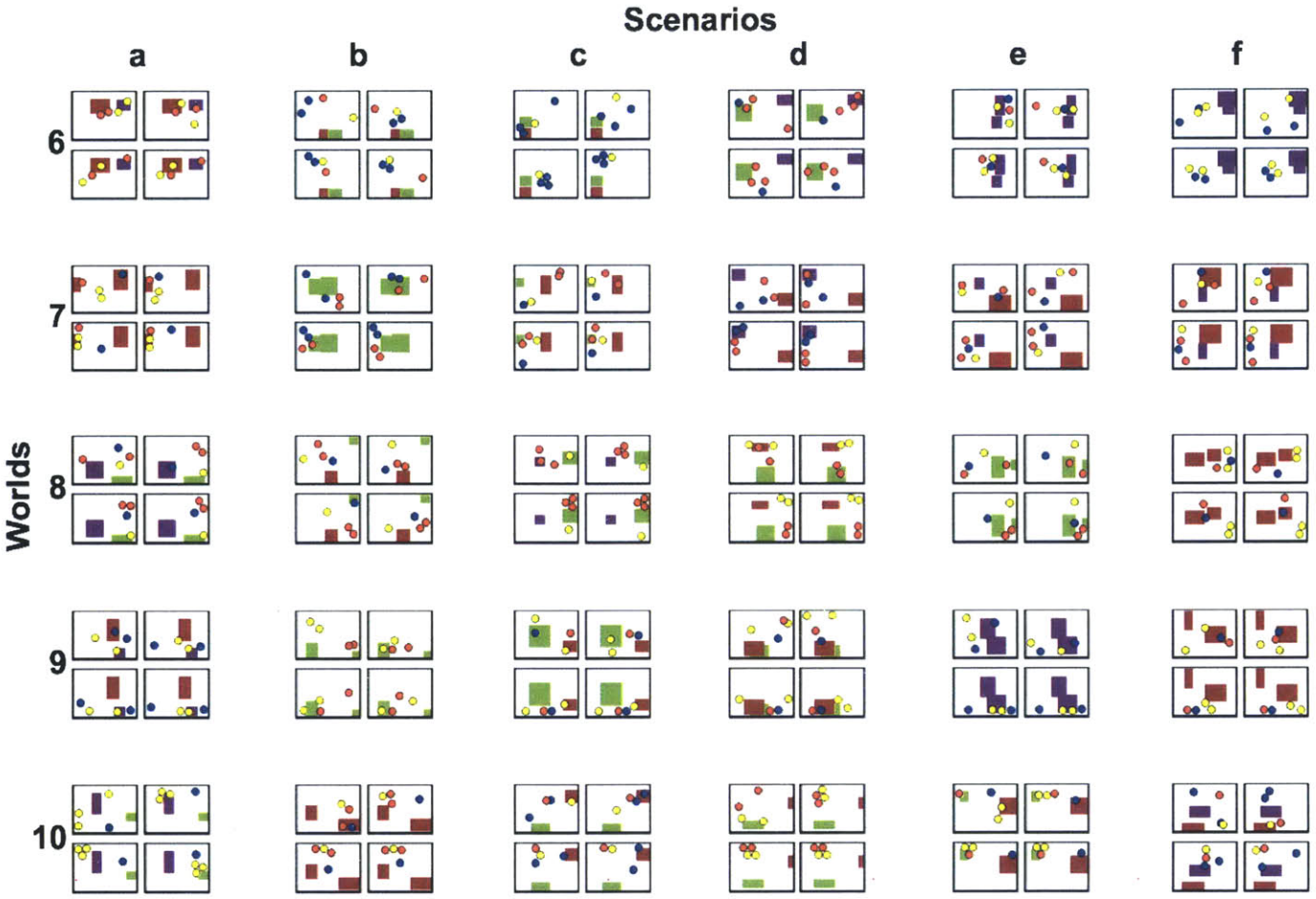


Figure 3-5: Part 2 of all the stimuli used, showing 'worlds' 6-10 with 6 scenarios per world. There are 4 images per scenario, showing it unfold over time. The images were sampled at the start each scenario (upper left image in each scenario), 1.25 seconds into the scenario (upper right image), 3.75 seconds into the scenario (lower left image) and at the end of the scenario (5 seconds after it started, lower right image).

with the complete instructions and layout used is available at `http://www.mit.edu/~tomeru/physics-experiment-turk/physics-experiment.html`.

After the presentation of each video participants rated the entire set of possible physical properties. For each puck color, participants were asked 'How massive are [color] objects?', with possible answers being 'Light', 'Medium', 'Heavy' or 'Can't tell from movie'. For each surface color, participants were asked 'How rough are [color] patches?', with possible answers being 'As smooth as the table-top', 'A little rough', 'Very rough' or 'Can't tell from movie'. For each puck color-pair combination, participants were asked 'How do [color 1] and [color 2] objects interact?', with possible answers being 'Attract', 'Repel', 'None', or 'Can't tell from movie'. Finally, participants were asked 'Is a global force pulling the objects, and if so in what direction is it pulling?', with possible answers being 'Yes, it pulls North','Yes, it pulls South', 'Yes, it pulls East', 'Yes, it pulls West' or 'No global force'. This gave us a total of 13 questions per video, and 5 videos gave us a total of 65 data points per participant. The 'Can't tell from video' answer was supplied for cases where the question is not relevant, for example a question regarding the mass of blue pucks when no blue pucks are shown in the video.

3.3.4 Results

Overview

Participants correctly answered 54% of the questions on average, with a standard error of 13%[5]. There was no statistically significant effect of learning over time (52% correct on first 2 videos vs. 55% answers on last 2 videos). This is far from

[5]The exact number of potentially correct questions varied by scenario, as some questions were not relevant for some stimuli, e.g. a question about the mass of blue pucks when no blue pucks were shown.

perfect, but we should not expect people to perform perfectly on a novel physical task. Rather, it is an accomplishment on the participants' part that they can adapt to a novel dynamical task at all. The participants' quantitative performance differed depending on the particular physical property being considered.

Analysis

We analyzed the results in two ways:

Aggregating over the different scenarios: We obtained the empirical distribution of responses over the possible answers across all scenarios. We collapsed across the property of color to consider four physical properties: mass, friction, pairwise forces and global forces. For mass and friction properties the responses were clearly ordinal (light, medium, and heavy for mass; smooth, a little rough, and very rough for friction) and the ground truth was a continuous ratio scale, thus we can fit an ordinal logistic regression to the participant data, shown in Fig. 3-6a. The figure displays the cumulative probability on the y-axis, and the relevant response is color-coded according to the label. For example, on this regression the probability people will answer 'light' when the true mass is in fact light (equal to 1) is 52%. The probability they will answer 'medium' is 33% (85%-52%), and the probability that they will answer 'heavy' is the remaining 15%. This is close to the empirical values of 47%/37%/16%.

An ordinal regression cannot be used for the global and coupling forces, and so Fig. 3-6c shows empirical confusion matrices, detailing the percentage of people that chose each option given the ground truth.

Transforming responses per scenario For mass and friction we can assess participant performance in a more refined way, by considering the distribution of re-

sponses for each puck (and surface) in each one of the 60 scenarios, and transforming this distribution into a quantitative prediction for that puck (or surface). We do this by taking the expectation of the physical property relative to the empirical distribution (e.g., if 60% of participants rated a yellow puck in scenario 7 as 'heavy' and 40% rated it as 'medium', the converted participant rating is $0.6 * 9 + 0.4 * 3 = 6.6$), and comparing the results with the ground truth, shown in Fig. 3-6b. These sub-figures plot the average rating of participants for mass/friction in a given scenario, compared to the 'ground truth'. Each black dot thus represents the average rating of 25-30 participants for mass/friction. The black solid line shows the average response for all masses across all scenarios. Dotted colored lines connect masses/friction in the same scenario, thus a rising line means a correct ranking. We next consider each property separately.

Results by physical property

Mass: The upward trend of the lines in the logistic regression, shown in Fig. 3-6a, shows that participants correctly shift in the probability of answering that a mass is heavier when that is in fact the case. The linear correlation depicted in Fig. 3-6b shows that although there is a large degree of variance for any given mass, participants were able to overall correctly scale the masses. The apparent ability to correctly rank and quantitatively scale multiple masses is of particular interest, as experiments on inferring mass from collisions have usually focused on judgments of mass ratios for two masses, often requiring binary responses of 'more/less massive' (e.g. [55]).

Friction: Again we see a upward trend in the logistic regression, shown in Fig. 3-6a. Compared with the regression for the masses, participants lean more heavily

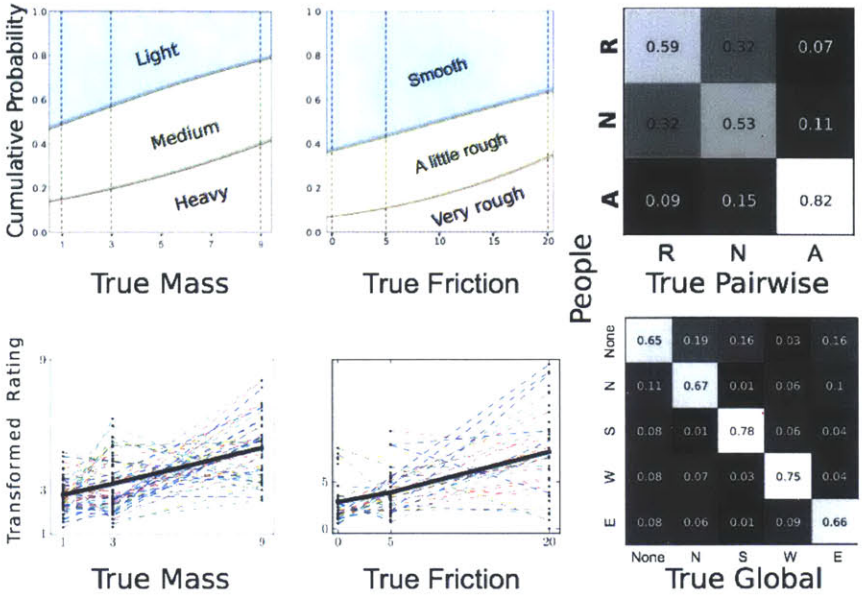


Figure 3-6: Analysis of participant performance using: (a) Ordinal logistic regression for mass (left) and friction (right). Shaded black areas represent uncertainty on parameter estimates, colored patches show the ordinal responses. The upward trend indicates a greater proportion of participants selecting the qualitatively correct response as the quantitative value goes up, (b) Per scenario analysis with transformed ratings for mass (left) and friction (right). Each black dot represents the average rating of 25-30 participants. The solid line shows the average response across all scenarios. Dotted lines connect mass/friction ratings in the same scenario, and so a rising line means a correct ranking. (c) Confusion matrices for pairwise forces (top) and global forces (bottom).

towards the lower end of the responses, perhaps because a 'null' response (no friction) is easier to make than a graded response along a continuum. The linear correlation depicted in Fig. 3-6b shows that participants were also able to correctly rank the roughness of the surfaces, though they could better distinguish between high- and low-friction surfaces than they were able to distinguish low- and zero-friction surfaces. To our knowledge this is the first systematic study of people's ranking of the friction properties of surfaces in the intuitive

physics literature.

Pairwise forces: As shown in Fig. 3-6c participants performed well on attraction forces, correctly detecting them on average in 82% of the cases in which they existed, while not reporting them on average in 88% of the cases in which they did not exist. As for repulsion and non-forces, their performance was above chance, although it was significantly worse than attraction. Note in particular that there is an asymmetry in the column for non-forces, indicating participants are confusing repulsion and non-existent forces, much more than they are confusing attraction and non-forces (32% vs. 15%). We will return to this point in the next section.

Global forces: As shown in Fig. 3-6c participants performed relatively well on detecting global forces, identifying the correct global force 70% of the time on average. Note that generally any force is more likely to be confused with a null-force than it is with any other force. Also, note that if participants did not correctly interpret the display as shown from a 'bird's eye view', then the 'South' direction could be interpreted as 'Down' and so activate certain prior expectations about a gravity force pulling in that direction. While this was indeed the most correctly perceived force, it is not a large effect, and such an explanation does not account for why a force pushing West, for example, is better detected than one pushing East.

3.4 Comparison to Ideal-Observer and Summary-Statistic Approximations

"Intuition" – A participant explaining how they arrived at their answers

For the *Ideal Observer* model (IO), we get predictions in the following way: For each scenario, we fix the observed initial conditions and simulate the resulting paths for all the relevant models. We then give each model a log-likelihood score by assessing the deviation of its simulated path from the observed path. Finally, for each parameter of interest we marginalize over the other parameters by summing them out, to obtain a log-likelihood score for the relevant parameter (see Fig. 3-3a and b).

For the *Simulation and Summary Statistics* model (SSS), we get predictions by following the procedure detailed at the end of Section 2. We also consider a simple combination of these two approaches, by summing weighted log-likelihoods from both approaches for any given physical parameter (IO&SSS) and renormalizing. These various approaches are illustrated for a particular example in Fig. 3-3.

These parameter estimates give us predicted distributions over the responses for each physical property for each scenario. We begin by collapsing across scenarios so that we can compare the results to the logistic regressions and confusion matrices of the participant data shown in Fig. 3-6a and c. Note that for each model there is a free 'noise' parameter applying to the distributions across all scenarios, which allows us to try and bring each model as close as possible to the participant data. We consider 'close' as minimizing the RMSE between the different distributions of the empirical confusion matrices (for pairwise and global forces) or the confusion matrices predicted by the logistic regression (for mass and friction)[6].

We begin by considering the ordinal logistic regression as applied to the different

[6]We also considered using KL-divergence as the distance metric, but that does not alter the results.

models, compared with mass and friction, shown in Fig. 3-6. For mass inference, the SSS model outperforms the IO model and is quite close to people's performance. The combined IO&SSS model places its entire weight on the SSS model, gaining no advantage from the IO model. For friction inference, we see that again the SSS model outperforms the IO model in terms of how close it is to people's judgments, although here the combined IO&SSS outperforms both.

We next consider the confusion matrices. Of particular interest is the confusion matrix for pairwise forces, where people showed an asymmetry in their confusion of the absence of force. That is, when there actually is an absence of a pairwise force, people incorrectly rate this as a repulsive force much more than they incorrectly rate this as an attractive force (32% repulsive compared with 15% for attractive, see Fig. 3-6c). We can understand this difference intuitively – an attractive force is more likely to pull bodies closer together, which makes the attraction stronger and so gives further evidence for the attractive force. A repulsive force pushes bodies further apart, growing weaker and providing less evidence for its existence over time. But such an asymmetry plays out over the entire dynamic scene. This asymmetry does not come naturally out of the IO model, which sums up the error along local deviations between a simulated trajectory given by a particular theory, and the observed trajectory. In such a model the local error produced by a theory that posits an attractive pairwise force is the same as that produced by a theory that posits a repulsive force.

"The only thing I would question is about the balls interaction. When they attract, that is easy enough to understand."
– A participant

By contrast, a summary statistic looking at the average pairwise distance does replicate this asymmetry. As illustrated in Fig. 3-3c, when we condition on the absence of force (in gray) and on a repulsive force (red), we generally find an overlap in the distribution of the summary statistic that is greater than that between the absence of force and an attractive force (green). Again it is important to note that

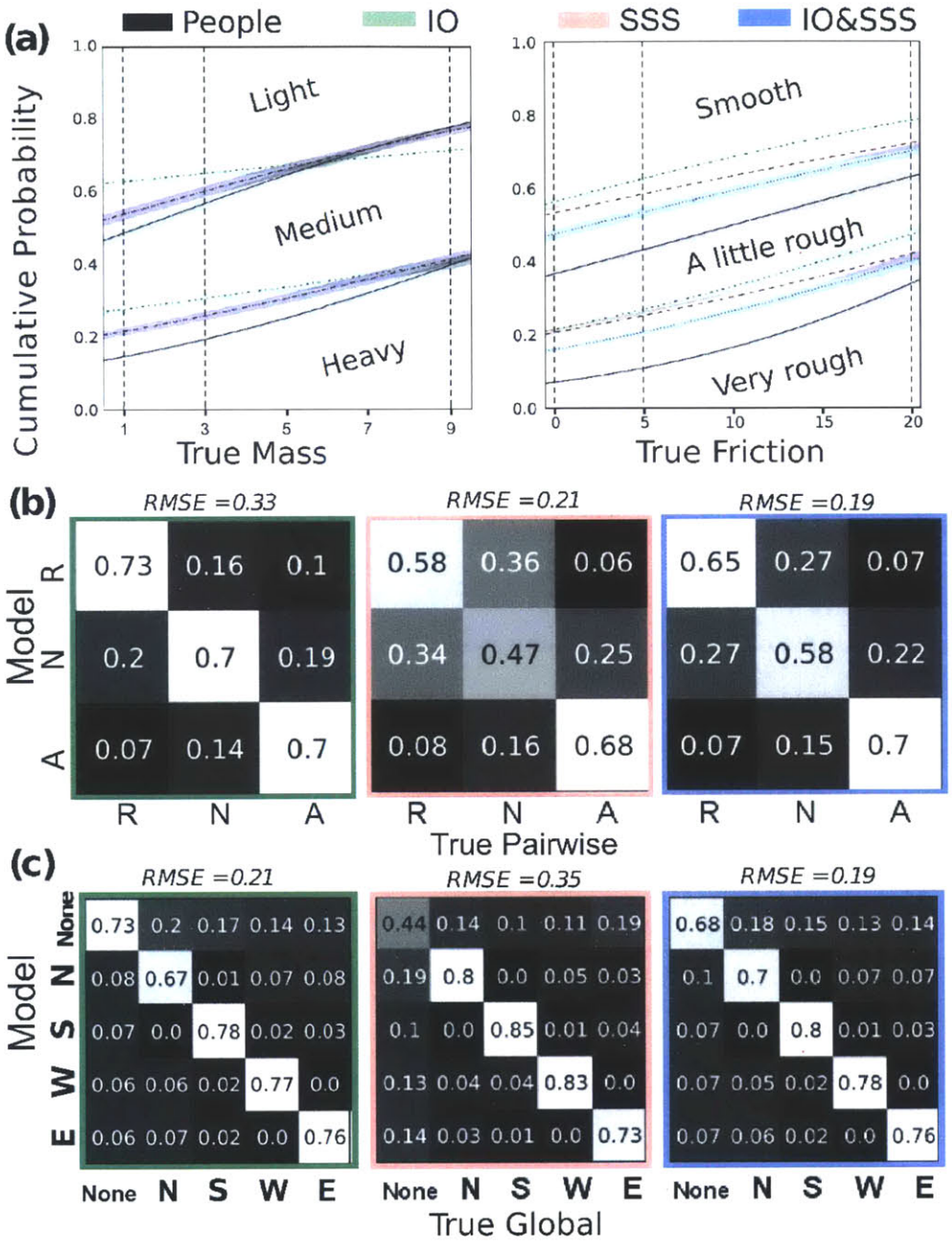


Figure 3-7: Comparison of model performance for properties (a) friction and mass (b) pairwise forces and (c) global forces.

the estimates from this summary statistic are informed by running many simulations using the formal model. When we combine the IO model with the SSS we can reproduce a confusion matrix that is similar to people's performance, shown in Fig. 3-6c. In particular, we reproduce the asymmetry between repulsion and the absence of a pairwise force (27% repulsive compared with 15% for attractive). While this asymmetry also exists for the SSS confusion matrix, the IO&SSS confusion matrix is closer to that of people.

The second confusion matrix to consider is that of global forces. As mentioned, for people one of the main points of interest was the confusion between any given force and the absence of force, relative to any other force. Both the IO and SSS models replicate this finding, although the IO model is in general closer to people. Also, we interestingly find that the SSS model is quite bad at detecting the absence of global forces, perhaps because none of the simple features we used account for a null-force. Again, a combination of the two into an IO&SSS produces a confusion matrix which is closest to that of people. We take up the question of other possible features, including more force-based ones, in the discussion.

Having examined the aggregate results, we can refine our comparison by looking at the response distributions the models give in each scenario and for each object and property, correlated with those of people. For mass and friction coefficient judgments, we can compare between people and the different approaches by again converting posteriors into predicted mass and friction values. For global and pairwise forces we can compare performance by correlating the predicted model posteriors for each scenario and property with the posterior as calculated from normalized people judgments.

The comparison of these various approaches with people is summarized in the table below, showing correlations between people and different approaches. Note that

we could theoretically have chosen the 'noise' parameter mentioned earlier to optimize this linear correlation, however we decided to reduce the number of free parameters and re-used the noise obtained from the previous comparison. We used a standard bootstrap method to obtain estimated confidence intervals on these correlations [36].

	Models		
	IO	SSS	IO&SSS
mass	0.50 ± 0.10	0.55 ± 0.08	0.55 ± 0.07
friction	0.54 ± 0.12	0.64 ± 0.11	0.65 ± 0.10
pairwise	0.56 ± 0.04	0.75 ± 0.03	0.81 ± 0.02
global	0.89 ± 0.02	0.85 ± 0.03	0.91 ± 0.02

Figure 3-8: Table showing the correlation between people's judgments of different physical properties and the different computational approaches: Ideal Observer (IO), Summary Statistics Approximation (SSS), and a combination of the two (IO&SSS). Correlations include 95% estimated confidence intervals, calculated using bootstrap methods.

As can be seen from the table, while not improving the results in all cases, the consistently best fit to people's judgments is obtained by using a combination of the ideal observer with simulation-based summary statistics methods. We show these correlations in more detail in Fig. 3-9. This suggests that a combination of the ideal observer with summary statistics discovered by generative simulations may be a future fruitful approach, an idea we take up in the general discussion.

3.5 General Discussion

Humans acquire their most basic physical concepts early in development, but continue to enrich and expand their intuitive physics throughout life as they are ex-

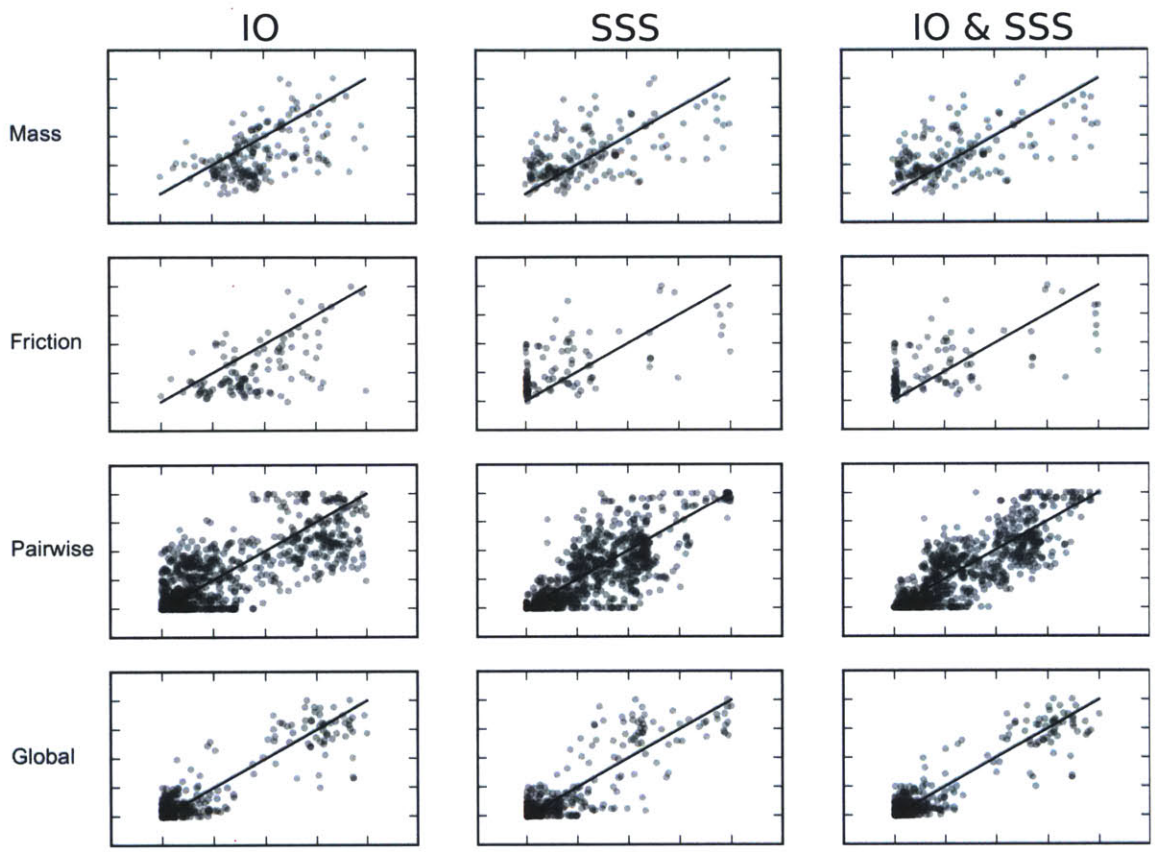


Figure 3-9: Correlations between people's answers and those given by the different models, for the four physical categories.

posed to more and varied dynamical environments. We have presented a hierarchical Bayesian framework to explain how physical theories can be learned across multiple timescales and levels of abstraction. Expressing theories using probabilistic programs lets our approach effectively learn the forces and properties that govern how objects interact in dynamic scenes unfolding over time. Given a challenging task of jointly inferring several novel physical laws from short movies through observation alone, people performed relatively well. Their performance was broadly in line with model predictions, but they also made systematic errors suggestive of how a bottom-up

summary-statistics-based approximate inference scheme might complement a more top-down ideal Bayesian approach to learning.

We found that on a number of measures, a hybrid between top-down Bayesian learning and bottom-up approximate inference emerged as the best empirical fit to participants' behavior in learning physical laws from dynamic scenes. This general approach also makes good engineering sense: It can transcend inherent limitations of each component method and serve as the basis for more robust real-world learning. The ideal Bayesian observer uses evidence in an optimal way, but it is computationally intractable. The feature-based statistics are useful heuristics in many cases, but are unable to handle situations that deviate from the norm [7]. Also, summary statistics in our setup do not replace the knowledge of a generative model, since they themselves require the simulations of a generative model to be computed. The computational intensity of the full ideal model is not as much of a problem in the combined model, as it is meant to capture either training the approximate, bottom-up inference in an off-line manner, or being used to score hypotheses once the bottom-up inference has narrowed the possible space down.

We considered a simple way of linearly combining the top-down and bottom-up models. While this approach performed reasonably, it does not get around the need to search a large space of theories for the ideal observer. A more psychologically plausible mechanism might include using the summary statistics of a given scenario to pick out a small space of 'reasonable' theories and then use Bayesian inference on this smaller space. For example, suppose the summary statistics of a scenario

[7] For example, consider a scenario involving two attracting pucks that begin in full contact, rotating around one another and moving together when one is struck. A normally useful statistic for detecting attraction - the difference between the initial and final distance of the pucks - would be useless here. The ideal observer and presumably people would have no problem detecting attraction in such a case

heavily bias in favor of an attractive force, less for the absence of force, and hardly at all for a repulsive force. A Bayesian inference mechanism with finite resources might then sample a handful of trajectories, most of them from theories that assume attractive forces, few from theories assuming no forces, and hardly any trajectories from theories assuming repulsive forces.

While we used of set of plausible summary statistics, it is not meant to be exhaustive. The fact that the Ideal Observer model performed better than the Summary Statistics Simulation model on some properties might be due to other unaccounted for features that, when used correctly, would bring the SSS model closer to people's performance for those properties as well. In particular, given the relation between forces and acceleration, it might be that more acceleration-based features would improve performance on force-related inference [8].

There are many questions that are still open when considering the challenge of inferring physical dynamics from perceptual scenes. In the rest of the discussion we consider several of these questions, and how our framework might shed light on them.

"A child who can catch a ball knows a good deal about trajectories"
– B.F. Skinner

First, to what extent are the computational processes underlying intuitive physics shared between adults and children? While it is clear that some physical knowledge develops [124, 3], it is possible that the highest level of the framework, such as an understanding of entities, forces and dynamics, is innate or early developing. Our own experiments focused on adults, but one advantage of our novel stimuli is that they can be easily adapted to experiments with young children or infants, using simple responses or violation of expectation to indicate what they learn from brief exposures.

[8] In order to facilitate the exploration of other features, the full participant responses as well as the trajectory data for all stimuli will be available at `http://www.mit.edu/~tomeru/physics2014/data/`

Second, how does the language people use to talk about physical properties relate to quantitative descriptions of those properties? In our task and in day-to-day physical descriptions we use words like "heavy" or "rough", which describe continuous qualities. These words are also graded adjectives with context-sensitive boundaries. An addition to our model could include drawing such properties from continuous distributions, such as different power-law distribution for the meaning of the words "light" and "heavy". We did not originally use such distributions because then even the ideal optimal inference model must be approximated, as the space of continuous concepts cannot be searched and scored exhaustively. Such an approximation raises questions about the exact technique to use, without allowing us to compare between ideal and approximate techniques, but it is possible and worth exploring[9].

The unintuitive trajectory of double pendulum

Third, what kind of physical forces, properties and dynamics do people find natural? What is intuitive in intuitive physics? In our framework we used pairwise and global forces, friction, collisions and stable conserved properties shared across objects, and people seemed able to reason about these relatively well. We believe people are able to reason about spring- and string-like forces, as well as attachments that maintain certain constraints on object relations. But it is entirely possible for our framework to generate and explore what we think will be non-intuitive dynamical scenes that people will find difficult to reason about, such as time-dependent, velocity-dependent forces that act according to non-conserved properties of objects. However, these forces would be more difficult to express in traditional physics simulations, suggesting a possible link to explore between simplicity in description length and human reasoning in intuitive physics.

Finally, what are the perceptual inputs that go into physical reasoning? Are they simply pixels that get grouped into 'motion features' used for bottom-up classifica-

[9]See for example [187] on approximate search in large theory spaces.

tion, or are the inputs properties of objects? This debate parallels the top-down vs. bottom-up questions of object recognition in visual perception, and like that debate it might turn out to not be an either-or distinction [184, 102]. Useful motion features might be real, but learned. Our framework suggests at least tentatively that new features for rapid classification might be partially discovered by using synthetic data which was generated by running forward many simulations from an intuitive physics model of the world, rather than relying on experience in the absence of such a model.

3.6 Conclusion

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

– Isaac Asimov

We have proposed that the combination of hierarchical Bayesian learning, an expressive representation for dynamical theories in terms of probabilistic programs, and psychologically plausible feature-based approximate inference schemes, offers a powerful framework for explaining how people can learn aspects of intuitive physics from observations - even such sparsely observed data as a few seconds of several objects in motion. Although participants were far from ideal observers in our experiments, they were nonetheless able to make inferences about all aspects of a given scenario's physics at levels well above chance, and these inferences could serve as important first steps guiding subsequent causal learning.

Much recent work on the development of intuitive theories has emphasized the crucial role that active interventions - and not only observational data - play in making causal learning possible. Likewise in science, experimental interventions -

and not simply correlational studies - have long been the gold standard for testing causal hypotheses. Yet controlled experiments and other interventions are not the only mode by which scientists and children learn about the world. They may not even be the most important. As Asimov suggests, every truly novel discovery in science begins with a moment of observation, a 'Thats funny...' moment, when a keen observer notices that something isn't quite as she expected, and differs from the usual course of events in a way that is not simply random but has some novel structure that calls for out exploration, experimentation and ultimately explanation.

We believe that this is just as true in the development of intuitive theories as in the development of formal scientific theories, and our studies here have aimed to capture this first step of learning in the domain of intuitive dynamics. In our experiments, the 'Thats funny...' moment might occur when two objects veer slightly off their straight-line course towards one another, or when an object slows down more than expected while moving over a colored surface. In our modeling, probabilistic programs express the knowledge by which people imagine how a scene might play out under different candidate physical laws or parameters, and how, if the scene departs from the imagined path, parts of the original program might be adjusted to account for the surprising data. These hypothetical adjustments become the hypotheses to be tested in subsequent experiments, and with luck, the seeds of "Eureka!".

3.7 Afterthought - Physics Engine Hacks for Psychology

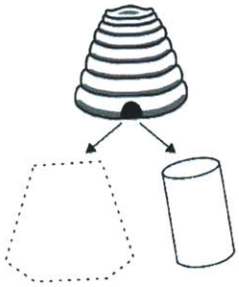
Physics engines do not fully simulate physics. Engineers, designers, physicists and computer scientists working on physics engines aren't concerned with getting a simu-

lation to perfectly account for the movement of each atom[10]. The modelers often use dynamics that are similar to Newtonian mechanics because that is the way the world works, but they're willing to also use shortcuts and hacks to get around problems of memory and computation.

If we take seriously the idea that people have something like a game-engine or physics-engine in their heads, then we should consider the concepts and workarounds that people working on physics engines have developed independently of psychology. Below I review a list of concepts that appear in many physics engines, and posit possible connections between them and concepts in cognitive science.

This is not to suggest that all the inner workings of physics engines will have counter-parts in the mind. But if engineers had to explicitly come up with clever ways to simulate the world around them, perhaps the mind uses similar ways. At the least, I hope this provides a fruitful avenue for future research.

Bodies and Shapes Many physics engines have a distinction between bodies and shapes. The 'body' holds the physical properties (mass, position, velocity, rotation, etc.), a bit like point-particles in physics, except that they can rotate. A body has one or more 'shapes' attached to it. These are the visible graphical bits. In 3D engines one can also find a distinction between 2 "meshes". Again, one is the actual 'physical' mesh, while the other is the the visible graphical one. Think of a bee-hive. As a graphical representation one can use some drawing or complicated mesh, but as a physical representation the engine will use some convex hull that envelops the graphical shape and allows for fast calculations, or possibly even a cylinder or pyramid. The simplified convex hull mesh, or the approximating cylinder is what is used for collision detection and physical



Simplifying a shape using a convex hull or cylindrical body

[10]Unless the physicists are trying to simulate atoms.

dynamics. It may be that when people simulate some object moving forward in time (say, throwing a bee-hive) they only roughly approximate that object using simpler meshes or bodies.

Dynamics and Collision Detection Most game and physics engines are split into dynamics (for moving things along) and collisions (for when things move into each other). Collisions seem fundamentally important, although they are detected and solved differently in different engines. There are many different hacks for noticing collisions (e.g. 'casting' trajectories geometrically into the future and seeing what they run into) and solving them (e.g. placing springs in between the colliding objects), but if physics engines exist in the mind, they will also have to work out the problem of collisions.

Static and Dynamic A common way to save on computation time and memory is to have a notion of "this body is not going to move", whether it is the background, the ground, a wall, etc. A static object is not just a very heavy object that you have to keep solving the forces and mass-reaction for, it is unmoving and does not participate in updating its own position properties. Such static entities might not count as 'Spelke-objects', and therefore violations of expectation tasks commonly associated with Spelke-objects would not apply to them.

"Walls are special" – Spelke (personal communication)

Sleeping and Awake While static bodies are those that are unlikely to move, physics engines also don't want to bother with dynamic bodies if possible. For example, if a dynamic body hasn't moved or contacted a body since the last frame, there is no point in graphically re-rendering it. A body "wakes up" when it collides with an awake body or has a joint destroyed. Psychologically,

this notion might explain certain attention effects.

Constraints and Joints These are things that restrict system of bodies without the need for explicit force simulation. Consider a two-bodied pulley system: The physics engine does not work through the exact tension on the rope in order to simulate a force that pulls one mass while the other goes down. Rather, there is a general constraint that "when one object moves up, the other moves down". Common constraints include distance joints, prismatic joints and revolution joints, but there are others. Again, such constraints seem psychologically useful, and they are in line with suggestions from 'qualitative physics' [45].

Fluids and Hard Things Fluids are a category onto themselves in most engines, and are trickier to simulate than single objects. There are many ways to approximate fluids, and there is probably an entire research program of trying to capture human reasoning about fluids by using different game engines. My main point in mentioning this category is that engines find this hard, and humans seem to find it hard as well. Stuff (fluids, sand piles, etc.) doesn't seem to obey 'Spelke object' principles [26, 83], but it might still be part of the physical reasoning system, and hard to reason about for the same reason engines have a hard time.

Chapter 4

Theory Learning as Stochastic Search*

If a person should say to you "I have toiled and not found", don't believe. If they say "I have not toiled but found", don't believe. If they say "I have toiled and found", believe. — Rabbi Itz'hak, Talmud

4.1 Introduction

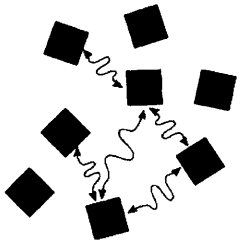
For the Rabbis of old, learning was toil, exhausting work – a lesson which many scientists also appreciate. Over recent decades, scientists have toiled hard trying to understand learning itself: what children know when, and how they come to know

*Joint work with Noah Goodman and Josh Tenenbaum

it. How do children go from sparse fragments of observed data to rich knowledge of the world? From one instance of a rabbit to all rabbits, from occasional stories and explanations about a few animals to an understanding of basic biology, from shiny objects that stick together to a grasp of magnetism – children seem to go far beyond the specific facts of experience to structured interpretations of the world.

What some scientists found in their toil is themselves. It has been argued that children's learning is much like a kind of science, both in terms of the knowledge children create, its form, content, and function, and the means by which they create it. Children organize their knowledge into *intuitive theories*, abstract coherent frameworks that guide inference and learning within particular domains [23, 26, 191, 64, 123]. Such theories allow children to generalize from given evidence to new examples, make predictions and plan effective interventions on the world. Children even construct and revise these intuitive theories using many of the same practices that scientists do [152]: searching for theories that best explain the data observed, trying to make sense of anomalies, exploring further and even designing new experiments that could produce informative data to resolve theoretical uncertainty, and then revising their hypotheses in light of the new data.

Consider the following concrete example of theory acquisition which we will return to frequently below. A child is given a bag of shiny, elongated, hard objects to play with, and finds that some pairs seem to exert mysterious forces on each other, pulling or pushing apart when they are brought near enough. These are magnets, but she doesn't know what that would mean. This is her first encounter with the domain. To make matters more interesting, and more like the situation of early scientists exploring the phenomena of magnetism in nature, suppose that all of the objects have an identical metallic appearance, but only some of them are magnetic, and only a subset of those are actually magnets (permanently magnetized). She may



initially be confused trying to figure out what interacts with what, but like a scientist developing a first theory, after enough exploration and experimentation, she might start to sort the objects into groups based on similar behaviors or similar functional properties. She might initially distinguish two groups, the magnetic objects (which can interact with each other) and the nonmagnetic ones (which do not interact). Perhaps then she will move on to subtler distinctions, noticing that this very simple theory doesn't predict everything she observes. She could distinguish three groups, separating the permanent magnets from the rest of the magnetic objects as well as from the nonmagnetic objects, and recognizing that there will only be an interaction if at least one of the two magnetic objects brought together is a permanent magnet. With more time to think and more careful observation, she might even come to discover the existence of magnetic poles and the laws by which they attract or repel when two magnets are brought into contact. These are but three of a large number of potential theories, varying in complexity and power, that a child could entertain to explain her observations and make predictions about unseen interactions in this domain.

Our goal here is to explore computational models for how children might acquire and revise an intuitive theory such as this, on the basis of domain experience. Any model of learning must address two kinds of questions: what, and how? Which representations can capture the form and content of *what* the learner comes to know, and which principles or mechanisms can explain *how* the learner comes to know it, moving from one state of knowledge to another in response to observed data? The main new contribution of this chapter addresses the 'how' question. We build on much recent work addressing the 'what' question, which proposes to represent the content of children's intuitive theories as probabilistic generative models defined over hierarchies of structured symbolic representations [177, 178, 90]. Previously the

'how' question has been addressed only at a very high level of abstraction, if at all: the principles of Bayesian inference explain how an ideal learner can successfully identify an appropriate theory, based on maximizing the posterior probability of a theory given data (as given by Bayes' rule). But Bayes' rule says nothing about the processes by which a learner could construct such a theory, or revise it in light of evidence. Here our goal is to address the 'how' of theory construction and revision at a more mechanistic, process level, exploring cognitively realistic learning algorithms. Put in terms of Marr's three levels of analysis [111], previous Bayesian accounts of theory acquisition have concentrated on the level of computational theory, while here we move to the algorithmic level of analysis, with the aim of giving a more plausible, practical and experimentally fertile view of children's developmental processes within the Bayesian paradigm.

Our work here aims to explain two challenges of theory acquisition in algorithmic terms. First is the problem of making learning work: getting the world right, as reliably as children do. As any scientist can tell you, reflecting on their own experiences of toil, the 'how' of theory construction and revision is nontrivial. The process is often slow, painful, a matter of starts and stops, random fits and bursts, missteps and retreats, punctuated by occasional moments of great insight, progress and satisfaction – the flashes of 'Aha!' and 'Eureka!'. And as any parent will tell you, children's cognitive development often seems to have much the same character. Different children make their way to adult-like intuitive theories at very different paces. Transitions between candidate theories often appear somewhat random and unpredictable at a local level, prone to backtracking or "two steps forward, one step back" behavior [158]. Yet in core domains of knowledge, and over long time scales, theory acquisition is remarkably successful and consistent: different children (at least within a common cultural context of shared experience) tend to converge

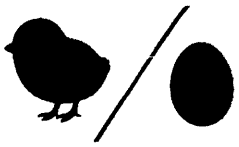
"I wonder why we think faster than we speak. Probably so we can think twice."
– Calvin and Hobbes

on the same knowledge structures, knowledge that is much closer to a veridical account of the world's causal structure than the infant's starting point, and they follow predictable trajectories along the way [26, 64, 190].

Our first contribution is an existence proof to show how this kind of learning could work – a model of how a search process with slow, fitful and often frustrating stochastic dynamics can still reliably get the world right, in part *because* of these dynamics, not simply in spite of them. The process may not look very algorithmic, in the sense of familiar deterministic algorithms such as those for long division, finding square roots, or sorting a list, or what cognitive scientists typically think of as a "learning algorithm", such as the backpropagation algorithm for training neural networks. Our model is based on a *Monte Carlo* algorithm, which makes a series of randomized (but not entirely random) choices as part of its execution. These choices guide how the learner explores the space of theories to find those that best explain the observed data – influenced by, but not determined by, the data and the learner's current knowledge state. We show that such a Monte Carlo exploratory search yields learning results and dynamics qualitatively similar to what we see in children's theory construction, for several illustrative cases.

Our second challenge is to address what could be called the "hard problem" of theory learning: learning a system of concepts that cannot be simply expressed as functions of observable sense data or previously available concepts – knowledge that is not simply an extension or addition to what was known before, but that represents a fundamentally new way to think. Developmental psychologists, most notably Susan Carey [26], have long viewed this problem of conceptual change or theory change as one of the central explanatory challenges in cognitive development. To illustrate, consider the concepts of "magnet" or "magnetic object" or "magnetic pole" in our scenario above, for a child first learning about them. There is no way to observe an

object on its own and decide if it falls under any of these concepts. There is no way to define or describe either "magnet" or "magnetic object" in purely sensory terms (terms that do not themselves refer to the laws and concepts of magnetism), nor to tell the difference between a "north" and a "south" magnetic pole from perception alone. How then could these notions arise? They could be introduced in the context of explanatory laws in a theory of magnetism, such as "Two objects will interact if both are magnetic and at least one is a magnet", or "Magnets have two poles, one of each type, and opposite types attract while like types repel." If we could independently identify the magnets and the magnetic objects, or the two poles of each magnetic object and their types, then these laws would generate predictions that could be tested on observable data. But only by virtue of these laws' predictions can magnets, magnetic objects, or magnetic poles even be identified or made meaningful. And how could one even formulate or understand one of these laws without already having the relevant concepts?



Theory learning thus presents children with a difficult joint inference task – a "chicken-and-egg" problem – of discovering two kinds of new knowledge, new concepts and new laws, which can only be made sense of in terms of each other: the laws are defined over the concepts, but the concepts only get their meaning from the roles they play in the laws. If learners do not begin with either the appropriate concepts or the appropriate laws, how can they end up acquiring both successfully? This is also essentially the challenge that philosophers have long studied of grounding meaning in *conceptual role* or *inferential role semantics* [14, 77, 78, 39, 41]. Traditional approaches to concept learning in psychology do not address this problem, nor do they even attempt to [20, 163, 137]. The elusiveness of a satisfying solution has led some scholars, most famously Jerry Fodor, to a radical skepticism on the prospects for learning genuinely new concepts, and a view that most concepts must

be innate in some nontrivial way [42, 43]. Carey [26] has proposed a set of informal "bootstrapping" mechanisms for how human learners could solve this problem, but no formal model of bootstrapping exists for theory learning, or concept learning in the context of acquiring novel theories.

We will argue that the chicken-and-egg problem can be solved by a rational learner but must be addressed in algorithmic terms to be truly satisfying: a purely computational-level analysis will always fail for the Fodorian skeptic, and will fail to make contact with the crux of the bootstrapping problem as Carey [26] frames it, since for the ideal learner the entire space of possible theories, laws and concepts, is in a sense already available from the start. An algorithmic implementation of that same ideal learning process can, however, introduce genuinely new concepts and laws in response to observed data. It can provide a concrete solution to the problem of how new concepts can be learned and can acquire meaning in a theory of inferential role semantics. Specifically, we show how a Monte Carlo search process defined over a hierarchically structured Bayesian model can effectively introduce new concepts as blank placeholders in the context of positing a new candidate explanatory law or extending an existing law. The new concept is not expressed in terms of pre-existing concepts or observable data; rather it is posited as part of a candidate explanation, together with pre-existing concepts, for observed data. In testing the candidate law's explanatory power, the new concepts are given a concrete interpretation specifying which entities they are most likely to apply to, assuming the law holds. If the new or modified law turns out to be useful – that is, if it leads to an improved account of the learner's observations, relative to their current theory – the law will tend to be retained, and with it, the new concept and its most likely concrete grounding.

"If concept learning is as Hypothesis Formation understands it, there can be no such thing."
– Fodor, LOT2

The rest of the chapter is organized as follows. We first present a nontechnical overview of the "what" and "how" of our approach to theory learning, and con-

trast it with the most well-known alternatives for modeling cognitive development based on connectionism and other emergentist paradigms. We then describe our approach more technically, culminating in a Markov Chain Monte Carlo (MCMC) search algorithm for exploring the space of candidate theories based on proposing random changes to a theory and accepting probabilistically those changes that tend to improve the theory. We highlight two features that make the dynamics of learning more efficient and reliable, as well as more cognitively plausible: a prior that proposes new theoretical laws drawn from *law templates*, biasing the search towards laws that express canonical patterns of explanation useful across many domains, and a process of *annealing* the search that reduces the amount of random exploration over time. We study the algorithm's behavior on two case studies of theory learning inspired by everyday cognitive domains: the taxonomic organization of object categories and properties, and a simplified version of magnetism. Finally, we explore the dynamics of learning that arise from the interaction between computational-level and algorithmic-level considerations: how theories change both as a function of the quantity and quality of the learner's observations, and as a function of the time course of the annealing-guided search process, which suggests promising directions for future experimental research on children's learning.

4.2 A nontechnical overview

A proposal for *what* children learn and a proposal for *how* they learn it may be logically independent in some sense, but the two are mutually constraining. Richer, more structured accounts of the form and content of children's knowledge tend to pose harder learning challenges, requiring learning algorithms that are more sophisticated and more costly to execute. As we explain below, our focus on explaining

the origins of children's intuitive theories leads us to adopt relatively rich abstract forms of knowledge representations, compared to alternative approaches to modeling cognitive development, such as connectionism. This leaves us with relatively harder learning challenges – connectionists might argue, prohibitively large. But we see these challenges as inevitable: Sooner or later, computational models of development must face them. Perhaps for the first time, we can now begin to see what their solution might look like, by bringing together recent ideas for modeling the form and content of theories as probabilistic generative models over hierarchies of symbolic representations [86, 87, 63] with tools for modeling the dynamics of learning as exploratory search based on stochastic Monte Carlo algorithms.

4.2.1 The 'What': Modeling the form and content of children's theories as hierarchical probabilistic models over structured representations

As a form of abstract knowledge, an intuitive theory is similar to the grammar of a language [176]: The concepts and laws of the theory can be used to generate explanations and predictions for an infinite (though constrained) set of phenomena in the theory's domain. We follow a long tradition in cognitive science and artificial intelligence of representing such knowledge in terms of compositional symbol systems, specifically predicate logic that can express a wide range of possible laws and concepts [42, 44, 141]. Embedding this symbolic description language in a hierarchical probabilistic generative model lets us bring to bear the powerful inductive learning machinery of Bayesian inference, at multiple levels of abstraction [72, 178].

Fig. 4-1 illustrates this framework. We assume a domain of cognition is given, comprised of one or more systems of entities and their relations, each of which gives

rise to some observed data. The learner's task is to build a theory of the domain: a set of abstract concepts and explanatory laws that explain the observed data for each system in that domain. The learner is assumed to have a hypothesis space of possible theories generated by (and constrained by) some "Universal Theory". We formalize this Universal Theory as a probabilistic generative grammar, essentially a probabilistic version of a language of thought [42]. Within this universal language, the learner constructs a specific theory that can be thought of as a more specific language for explaining the phenomena of the given domain.

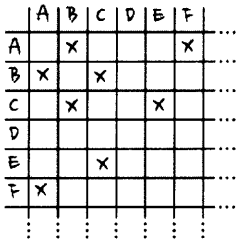
In principle, an ideal learner should consider all possible theories expressible in the language of thought and weigh them against each other in light of observed evidence. In practice, there are infinitely many candidate theories and it will be impossible to explicitly consider even a small fraction of them. Explaining how a learner proposes specific candidate theories for evaluation is a task for our algorithmic-level account (see below under 'How').

Candidate theories are evaluated using Bayes' rule to assess how likely they are to have generated the observed data. Bayes' rule scores theories based on the product of their prior probabilities and their likelihoods. The prior reflects the probability of generating the laws and concepts of a theory a priori from the generative grammar, independent of any data to be explained. The likelihood measures the probability of generating the observed data given the theory, independent of the theory's plausibility. Occam's razor-like considerations emerge naturally from a Bayesian analysis: the prior will be highest for the simplest theories, whose laws can be generated with the fewest number of a priori stipulations, while the likelihood will be highest for theories whose laws allow a domain to be described accurately and compactly, generating the observed data with a spare set of minimal facts.

The fit of a theory to data cannot be evaluated directly; its laws express the

abstract principles underlying a domain but no specific expectations about what is true or false. One level below the theory in the hierarchical framework, the learner posits a *logical model* of each observed system in the domain. The logical model, or "model" for short, specifies what is true of the entities in a particular system in ways consistent with and constrained by the theory's abstract laws. Each model can be thought of as one particular concrete instantiation of the abstract theory. It generates a probability distribution over possible observations for the corresponding system, and it can be scored directly in terms of how well those predictions fit the actual data observed.

As a concrete example of this framework, consider again the child learning about the domain of magnetism. She might begin by playing with a few pieces of metal and notice that some of the objects interact, exerting strange pulling or pushing forces on each other. She could describe the data directly, as "Object *a* interacts with object *j*", "Object *i* interacts with object *j*", and so on. Or she could form a simple theory, in terms of abstract concepts such as *magnet*, *magnetic object* and *non-magnetic object*, and laws such as "*Magnets* interact with other *magnets*", "*Magnets* interact with *magnetic objects*", and "Interactions are symmetric". It is important to note that terms like *magnet* convey no actual information about the object, and they are simply labels. Systems in this domain correspond to specific subsets of objects, such as the set of objects *a, ..., i* in Fig. 4-1. A model of a system specifies the minimal facts needed to apply the abstract theory to the system, in this case which objects are magnetic, which are magnets, and which are non-magnetic. From these core facts the laws of the theory determine all other true facts – in our example, this means all the pairwise interactions between the objects: e.g., objects *i* and *j*, being magnets, should interact, but *i* and *e* should not, because the latter is non-magnetic. Finally, the true facts generate the actual data observed by the learner via a noisy



A table tallying interactions cannot generalize or compress data

sampling process, e.g. observing a random subset of the object pairs that interact, and occasionally misperceiving an object's identity or the nature of an interaction.

While the abstract concepts in this simplified magnetism theory are attributes of objects, more complex relations are possible. Consider for example a theory of taxonomy, as in Collins and Quillian's classic model of semantic memory as an inheritance hierarchy [31]. Here the abstract concepts are *is_a* relations between categories and *has_a* relations between categories and properties. The theory underlying taxonomy has two basic laws: "The *is_a* relation is transitive" and "The *has_a* relation inherits down *is_a* relations" (laws 3 and 4 on the "Taxonomy" column of Fig. 4-1). A system consists of a specific set of categories and properties, such as *salmon*, *eagle*, *breathes*, *can fly*, and so on. A model specifies the minimal *is_a* and *has_a* relations, typically corresponding to a tree of *is_a* relations between categories with properties attached by *has_a* relations at the broadest category they hold for: e.g., "A canary is a bird", "A bird is an animal", "An animal can breathe", and so on. The laws then determine that properties inherit down chains of *is_a* relations to generate many other true facts that can potentially be observed, e.g., "A canary can breathe".

The analogy between learning a theory for a domain and learning a grammar for a natural language thus extends down through all levels of the hierarchy of Fig. 4-1. A logical model for a system of observed entities and relations can be thought of as a parse of that system under the grammar of the theory, just as the theory itself can be thought of as a parse of a whole domain under the grammar of the universal theory. In our hierarchical Bayesian framework, theory learning is the problem of searching jointly for the theory of a domain and models of each observed system in that domain that together best parse all the observed data.[1]

[1]The idea of hierarchical Bayesian grammar induction, where the prior on grammars is itself generated by a grammar (or "grammar grammar"), dates back at least to the seminal work of

Previous applications of grammar-based hierarchical Bayesian models have shown how, given sufficient evidence and a suitable theory grammar, an ideal Bayesian learner can identify appropriate theories in domains such as causality [72, 63], kinship and other social structures [87], and intuitive biology [176]. While our focus in this chapter is the algorithmic level – the dynamics of how learners can search through a space of theories – we have found that endowing our theory grammars with one innovation greatly improves their algorithmic tractability. We make the grammar more likely to generate theories with useful laws by equipping it with law templates, or forms of laws that capture canonical patterns of coherent explanation arising in many domains. For example, law templates might suggest explanations for when an observed relation $r(X, Y)$ holds between entities X and Y (e.g., X attracts Y, X activates Y, X has Y) in terms of latent attributes of the objects, $f(X)$ and $g(Y)$, or in terms of some other relation $s(X, Y)$ that holds between them, or some combination thereof: perhaps $r(X, Y)$ holds if $f(X)$ and $s(X, Y)$ are both true. Explanatory chains introducing novel objects are also included among the templates: perhaps $r(X, Y)$ holds if there exists a Z such that $s(X, Z)$ and $s(Z, Y)$ hold. As we explain below, making these templates explicit in the grammar makes learning both more cognitively plausible and much faster.

The most familiar computational alternative to structured Bayesian accounts of cognitive development are connectionist models, and other *emergentist* approaches [115]. Instead of representing children's abstract knowledge in terms of explicit symbol systems, these approaches attribute abstract knowledge to children only implicitly as an 'emergent' phenomenon that arises in a graded fashion from interactions among more concrete, lower-level non-symbolic elements – often inspired loosely by neuroscience. *Dynamical systems* models view the nervous system as a complex adap-

Feldman and colleagues [38].

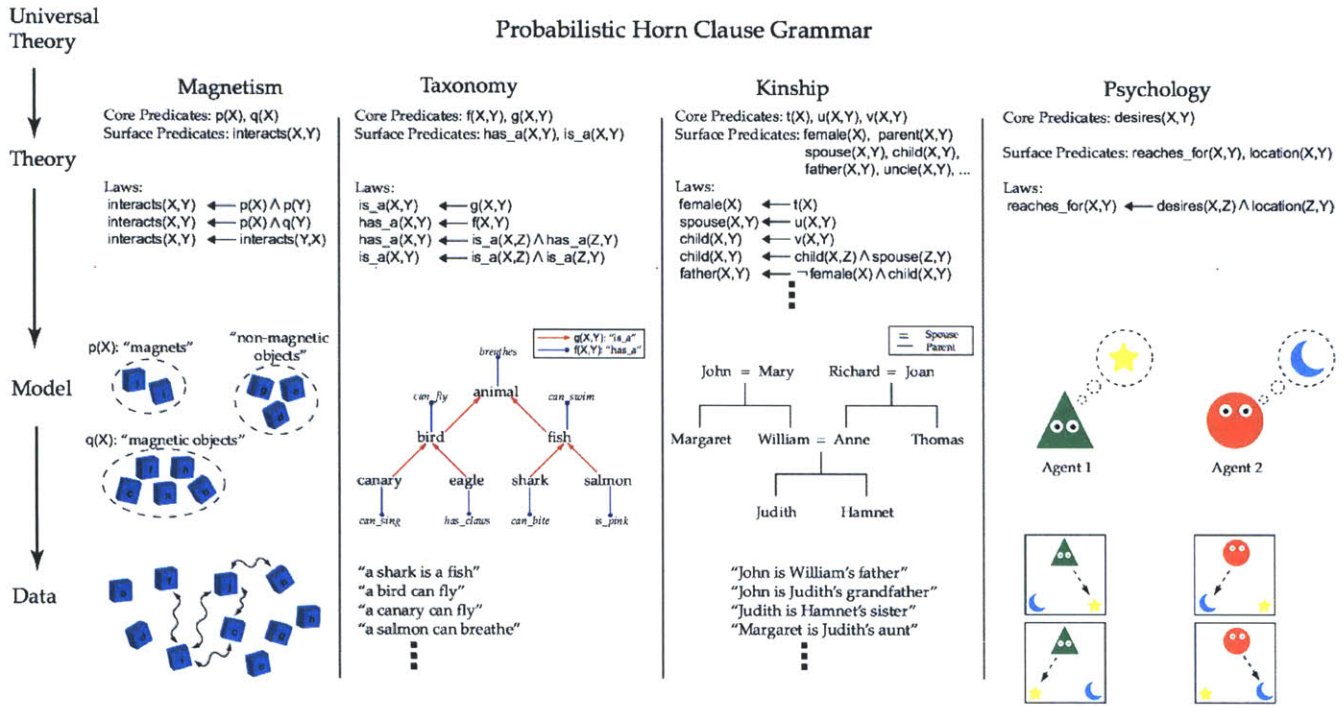


Figure 4-1: A hierarchical Bayesian framework for theory acquisition. Each level generates the space of possibilities for the level below, providing constraints for inference. Four examples of possible domain theories are given in separate columns, while the rows correspond to different levels of the hierarchy. A domain theory aims to explain observable values of one or more surface predicates by positing one or more core predicates and a set of simple laws relating them (perhaps supplemented by some background knowledge, as with the *location* predicate in the right-most column). The core predicates represent the minimal facts necessary to explain the observations; a model of a theory is then a particular extension of the core predicates to the objects in the domain. The observations are assumed to be a random sample of all the true facts given by the model. Probabilistic inference on this hierarchical model then supports multiple functions, including learning a theory from observed data, using a theory to derive the most compact model that explains a set of observations, and using that model to predict unobserved data.

tive system evolving on multiple timescales, with emergent behavior in its dynamics.

"Adjust the parameters of the mind in proportion to the extent to which their adjustment can produce a reduction...between expected and observed events" – McClleland

Connectionist models view children's knowledge as embedded in the strengths of connections between many neuron-like processing units, and treat development as the tuning of these strengths via some experience-dependent adjustment rule. Connectionists typically deny that the basic units of traditional knowledge representation – objects, concepts, predicates, relations, propositions, rules and other symbolic abstractions – are appropriate for characterizing children's understanding of the world, except insofar as they emerge as approximate higher-level descriptions for the behavior dictated by a network's weights.

While emergentist models have been well-received in some areas of development, such as the study of motor and action systems [115], emergentist models of the structure and origins of abstract knowledge [137] have not been widely embraced by developmentalists studying children's theories [64, 26]. There is every reason to believe that explicit symbolic structure is just as important for children's intuitive theories as for scientists' more formal theories – that children, like scientists, cannot adequately represent the underlying structure of a domain such as physics, psychology or biology simply with a matrix of weights in a network that maps a given set of inputs to a given set of outputs. Children require explicit representations of abstract concepts and laws in order to talk about their knowledge in natural language, and to change and grow their knowledge through talking with others; to reason causally in order to plan for the future, explain the past, or imagine hypothetical situations; to apply their knowledge in novel settings to solve problems that they have never before encountered; and to compose abstractions recursively, as in forming beliefs about others' beliefs about the physical world and how those beliefs might be different than one's own.

Despite these limitations, connectionist models have been appealing to developmentalists who emphasize the processes and dynamics of learning more than the nature of children's knowledge representations [156, 115]. This appeal may come from the fact that when we turn from the 'what' to the 'how' of children's learning, connectionist models have a decided advantage: learning in connectionist systems appears much better suited to practical algorithmic formulation, and much more tractable, relative to structured probabilistic models or any explicitly symbolic approach. As we explain below, making the 'how' of learning plausible and tractable may be the biggest challenge facing the structured probabilistic approach.

4.2.2 The 'How': Modeling the dynamics of children's theory learning as stochastic (Monte Carlo) exploratory search

It is helpful to imagine the problem children face in learning as that of moving over a "knowledge landscape", where each point represents a possible state of knowledge and the height of that point reflects the value of that knowledge-state – how well it allows the child to explain, predict, and act on their world. Such a picture is useful in showing some of the differences between our approach to cognitive development and the connectionist and emergentist alternatives, and it highlights the much more serious 'how' challenge that confronts structured probabilistic models.

Viewed in landscape terms (Fig. 4-2), connectionist models typically posit that children's knowledge landscape is continuous and smooth, and this matters greatly for the mechanisms and dynamics of learning. Learning consists of traversing a high-dimensional real-valued "weight space", where each dimension corresponds to the strength of one connection in a neural network. Fig. 4-2 depicts only a two-

dimensional slice of the much higher dimensional landscape corresponding to the three-layer network shown. The height of the landscape assigned to each point in weight space – each joint setting of all the network's weights – measures how well the network explains observed data in terms of an error or energy function, such as a sum-of-squared-error expression. The topology of these landscapes is simple and uniform: at any point of the space, one can always move along any dimension independently of every other, and changing one parameter has no effect on any other. The geometry is also straightforward: neighboring states, separated by small changes in the weights or parameters, typically yield networks with very similar input-output functionality. Thus a small move in any direction typically leads to only a small rise or fall in the error or energy function.

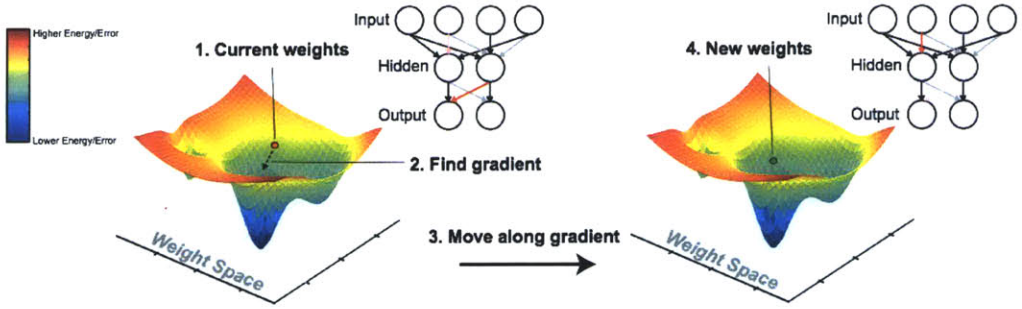


Figure 4-2: A hypothetical neural network and a weight space spanning the possible values of two particular connections. Steps 1-4 show the sequence of a learning algorithm in such a space: the calculation of a gradient and the move to a lower point. This corresponds to a shift in the network's connection weights and a smaller error on the output.

This geometry directly translates into the dynamics of learning: the Hebb rule, the Delta Rule, Backpropagation and other standard weight-adjustment rules [116] can be seen as implementing gradient descent – descending the error or energy landscape by taking small steps along the steepest direction – and it can be proven that

this dynamic reliably takes the network to a local minimum of error, or a locally best fitting state of knowledge. In certain cases, particularly of interest in contemporary machine learning systems [13], the error landscape can be designed to have a geometric property known as convexity, ensuring that any local minimum is also a global minimum and thus that the best possible learning end-state can be achieved using only local weight-adjustment rules based on gradient descent. Thus learning becomes essentially a matter of "rolling downhill", and is just as simple. Even in cases where there are multiple distinct local minima, connectionist learning can still draw on a powerful toolkit of optimization methods that exploit the fact that the landscape is continuous and smooth to make learning relatively fast, reliable and automatic.

Now consider the landscape of theory learning from the perspective of our structured Bayesian approach, and it should become clear how much more difficult the problem is (Fig. 4-3). Each point on the landscape now represents a candidate domain theory expressed in terms of one or more laws in first-order logic and one or more abstract concepts indicated by a blank predicate (e.g., $f(X)$, $g(X)$). Two possibilities for a simple theory of magnetism are shown, labeled Theory B and Theory C (these will be explained in much greater detail below). The height of the surface at a given point represents how well the corresponding theory is supported by the observed data, which we measure as the Bayesian posterior probability. (Note that in contrast to Fig. 4-2, where "lower is better", here "higher is better", and the goal is to seek out maxima of the landscape, not minima.) Unlike the weight space shown in Fig. 4-2, this portrait of a "theory space" as two-dimensional is only metaphorical: it is not simply a lower-dimensional slice of a higher-dimensional space. The space of theories in a language of thought is infinite and combinatorially structured with a neighborhood structure that is impossible to visualize faithfully on a page.

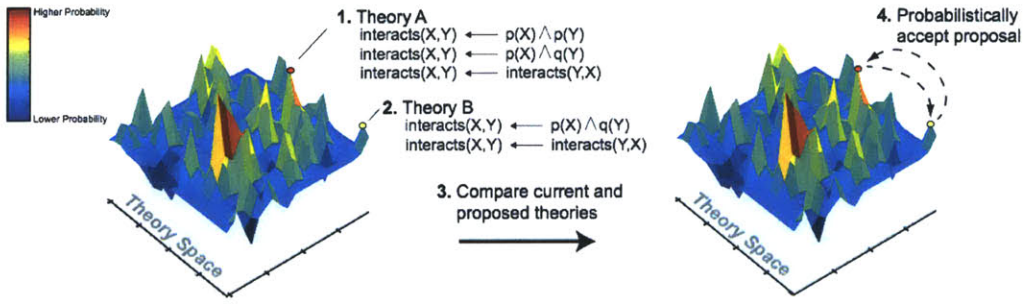


Figure 4-3: Schematic representation of the learning landscape within the domain of simple magnetism. Steps 1-4 illustrate the algorithmic process in this framework. The actual space of of theories is discrete, multidimensional and not necessarily locally connected.

At the level of computational theory, we can imagine an ideal Bayesian learner who computes the full posterior probability distribution over all possible theories, that is, who grasps this entire landscape and assesses its height at all points in parallel, conditioned on any given observed data set. But this is clearly unrealistic as a starting point for algorithmic accounts of children's learning, or any practical learning system with limited processing resources. Intuition suggests that children may simultaneously consider no more than a handful of candidate theories in their active thought, and developmentalists typically speak of the child's current theory as if, as in connectionist models, the learner's knowledge state corresponds to just a single point on the landscape rather than the whole surface or posterior distribution. The ideal Bayesian learner is in a sense similar to a person who has "not toiled but found" from the opening epigraph: the entire hypothesis space is already defined and the learner's task is merely to reshuffle probability over that space in response to evidence. The actual child must toil and construct her abstract theory, piece by piece, generalizing from experience.

Considering how a learner could move around on this landscape in search of the

best theory, we see that most of the appealing properties of connectionist knowledge landscapes – the features that support efficient learning algorithms – are not present here. The geometry of the landscape is far from smooth: A small change in one of the concepts or laws of a theory will often lead to a drastic rise or fall in its plausibility, leading to a proliferation of isolated local maxima. There is typically no local information (such as a gradient) diagnostic of the most valuable directions in which to move. The landscape is even more irregular in ways that are not easily visualized. There is no uniform topology or neighborhood structure: the number and nature of variants that can be proposed by making local changes to the learner's current hypothesis vary greatly over the space, depending on the form of that hypothesis. Often changing one aspect of a theory requires others to be changed simultaneously in order to preserve coherence: for instance, if we posit a new abstract concept in our theory, such as the notion of a *magnet*, or if we remove a conceptual distinction (such as the distinction between *magnets* and *magnetic objects*), then one or more laws of the theory will need to be added, removed or redefined at the same time.

Artificial intelligence has a long history of treating learning in terms of search through a discrete space of symbolic descriptions, and a wide variety of search algorithms have been proposed to solve problems such as rule discovery, concept learning and generalization, scientific law discovery, and causal learning [125, 121, 18, 127, 171]. For some of these problems, there exist systematic search algorithms that can be as fast and reliable as the gradient-based optimization methods used in connectionist learning [121, 127, 171]. But for problems like scientific discovery [18], or our formulation of children's theory learning, the best known search algorithms are not like this. Much like child learners, we suggest, these algorithms are slow, unreliable, and unsystematic (indeed often random), but with enough patience they can be expected to converge on veridical theories.

The specific search algorithm we describe is based on widely used methods in statistics and AI for approximating intractable Bayesian inferences, known as Markov Chain Monte Carlo (MCMC). MCMC algorithms have recently been proposed as models for the short-timescale dynamics of perceptual inferences in the brain [53, 173, 122], but they are also well-suited to understanding the much longer-term dynamics of learning.

The remainder of this section sketches how our MCMC algorithm answers the two main challenges we set out at the start of this chapter: explaining how children can reliably converge on veridical theories, given their constrained cognitive resources and a learning dynamic that often appears more random than systematic, and explaining how children can solve the hard "chicken-and-egg" inference problem of jointly learning new concepts and new laws defined in terms of those concepts.

The heart of MCMC theory learning is an iterative loop of several basic steps, shown in Fig. 4-3. The learner begins at some point in the theory landscape (e.g. theory B or C in Fig. 4-3). The learner then proposes a possible move to a different theory, based on modifying the current theory's form: adding/deleting a law or set of laws, changing parts of a law or introducing a new concept, and so on. The proposed and current theories are compared based on evaluating (approximately) how well they explain the observed data (i.e., comparing the relative heights of these two points on the theory landscape). If the proposed theory scores higher, the learner accepts it and moves to this new location. If the proposal scores lower, the learner may still accept it or reject it (staying at the same location), with probability proportional to the relative scores of the two theories. These steps are then repeated with a new proposal based on the new current location.

From the standpoint of MCMC, randomness is not a problem but rather an essential tool for exploring the theory landscape. Because MCMC algorithms consider

only one hypothesis at a time and propose local modifications to it, and there are no generally available signals (analogous to the error gradient in connectionist learning) for how to choose the best modification of the current hypothesis out of an infinite number of possible variations, the best learners can do is to propose variant theories to explore chosen in a randomized but hopefully intelligent fashion. Our algorithm proposes variants to the current hypothesis by replacing a randomly chosen part of the theory with another random draw from the probabilistic generative grammar for theories (that is, the prior over theories). This process could in principle propose any theory as a variant on any other, but it is naturally biased towards candidates that are most similar to the current hypothesis, as well as those that are a priori simpler and more readily generated by the grammar's templates for coherent laws. The use of law templates is crucial in focusing the random proposal mechanism on the most promising candidates. Without templates, all of the laws proposed could still have been generated from a more general grammar, but they would be much less likely a priori; learners would end up wasting most of their computational effort considering simple but uesless candidate laws. The templates make it likely that any random proposal is at least a plausibly useful explanation, not just a syntactically well-formed expression in the language of thought.

The decision of whether to accept or reject a proposed theory change is also made in a randomized but intelligently biased fashion. If a proposed change improves the theory's account of the data, it is always accepted, but sometimes a change that makes the theory worse could also be accepted. This probabilistic acceptance rule helps keep the leaner from becoming trapped for too long in poor local maxima of the theory landscape [57].

Although we use MCMC as a search algorithm, aiming to find the best theory, the algorithm's proper function is not to find a single optimal theory but rather

to visit all theories with probability proportional to their posterior probability. We can interpolate between MCMC as a posterior inference technique and MCMC as a search algorithm by *annealing* – or starting with more stochastic (or noisy) search moves and "lowering the temperature", making the search more deterministic over time [97, 166]. This greatly improves convergence to the true theory. Such an algorithm can begin with little or no knowledge of a domain and, given enough time and sufficient data, reliably converge on the correct theory or at least some approximation thereof, corresponding to a small set of abstract predicates and laws.

Annealing is also responsible for giving the MCMC search algorithm some of its psychologically plausible dynamics. It gives rise to an early high-temperature exploration period characterized by a large number of proposed theories, most of which are far from veridical. As we see in young children, new theories are quick to be adopted and just as quick to be discarded. As the temperature is decreased, partially correct theories become more entrenched, it becomes rarer for learners to propose and accept large changes to their theories, and the variance between different theory learners goes down. As with older children, rational learners at the later stages of an annealed MCMC search tend to mostly agree on what is right, even if their theories are not perfect. Without annealing, MCMC dynamics at a constant temperature could result in a learner who is either too conservative (at low temperature) or too aggressive (at high temperature) in pursuing new hypotheses – that is, a learner who is prone to converge too early on a less-than-ideal theory, or to never converge at all.

The idea of annealing comes from metallurgy: "The smith dips [the axe] in cool water to temper it, strengthening the iron" (The Odyssey)

Figures 4-6a and 4-7a illustrate these learning dynamics. (these are explained in detail in the next sections). On average, learners are consistently improving. On average, they are improving gradually. But individually, learners often get worse before they get better. Individually, they adopt theories in discrete jumps, signifying moments of fortuitous discovery. Such dynamics on the level of the individual learner

are more in line with discovery processes in science and childhood than are the smoother dynamics of gradient descent on a typical connectionist energy landscape. Critics might reasonably complain that MCMC methods are slow and unreliable by comparison. But theory construction just is a difficult, time-consuming, painful and frustrating business – in both science and children's cognition. We can no more expect the dynamics of children's learning to follow the much tamer dynamics of gradient learning algorithms than we could expect to replace scientists with a gradient-based learning machine and see the discoveries of new concepts and new scientific laws emerging automatically. [2] Currently we have no good alternative to symbolic representational machinery for capturing intuitive theories, and no good alternative to stochastic search algorithms for finding good points in the landscape of these symbolic theories.

What of the "hard problem of theory learning", the challenge of jointly learning new laws and new concepts defined only in terms of each other? Our MCMC search unfolds in parallel over two levels of abstraction: an outer loop in the space of theories, defined by sets of abstract laws; and an inner loop in the space of explanations or models generated by the theory for a particular domain of data, defined by groundings of the theory's concepts on the specific entities of the domain. This two-level search lets us address the "chicken and egg" challenge by first proposing new laws or changes to existing laws of a theory in the outer search loop; these new laws can posit novel but 'blank' concepts of a certain form, whose meaning is

וּמַרְאֵיהֶם וּמַעֲשֵׂיהֶם כְּאֲשֶׁר יִהְיֶה הָאוֹפַן בְּתוֹךְ הָאוֹפָן
-- יחזקאל א׳ טז׳

[2] It is worth noting that not all connectionist architectures and learning procedures are confined to gradient-based methods operating on fixed parametric architectures. In particular the constructivist neural networks explored by Tom Shultz and colleagues [156] are motivated by some of the same considerations that we are, aiming to capture the dynamics of children's discovery with learning rules that implement a kind of exploratory search. These models are still limited in their representational power, however: they can only express knowledge whose form and content fits into the connections of a neural network, and not the abstract concepts and laws that constitute an intuitive theory. For that reason we favor the more explicitly symbolic approach described here.

then filled in the most plausible way on the next inner search loop. For example, the algorithm may posit a new rule never before considered, that objects of type f interact with objects type g, without yet specifying what these concepts mean; they are just represented with blank predicates $f(X)$ and $g(X)$. The inner loop would then search for a reasonable assignment of objects to these classes – values for $f(X)$ and $g(X)$, for each object X – grounding them out as magnets and magnetic objects, for example. If this law proves useful in explaining the learner's observations, it is likely to persist in the MCMC dynamics, and with it, the novel concepts that began as blank symbols f and g but have now effectively become what we call "magnets" and "magnetic objects".

In sum, we see many reasons to think that stochastic search in a language of thought, with candidate theories generated by a probabilistic generative grammar and scored against observations in a hierarchical Bayesian framework, provides a better account of children's theory acquisition than alternative computational paradigms for modeling development such as connectionism. Yet there are also major gaps: scientists and young children alike are smarter, more active, more deliberate and driven explorers of both their theories and their experiences and experiments than are our MCMC algorithms [151]. We now turn to a more technical treatment of our model but we return to these gaps in the general discussion below.

4.3 Formal framework

This section gives a more formal treatment of theory learning, beginning with our hierarchical Bayesian framework for describing 'what' is learned (Fig. 4-8), and then moving to our proposed MCMC algorithm for explaining 'how' it could be learned (Fig. 4-3).

Formally, the hierarchical picture of knowledge shown in Fig. 4-8 provides the backbone for a multilevel probabilistic generative model: conditional probability distributions that link knowledge at different levels of abstraction, supporting inference at any level(s) conditioned on knowledge or observations at other levels. For instance, given a domain theory T and a set of noisy, sparse observations D, a learner can infer the most likely model M and use that knowledge to predict other facts not yet directly observed [86, 87]. The theory T sets the hypothesis space and priors for the model M, while the data D determine the model's likelihood, and Bayes' rule combines these two factors into a model's posterior probability score,

$$P(M|D,T) \propto P(D|M)P(M|T). \tag{4.1}$$

If the theory T is unknown, the learner considers a hypothesis space of candidate theories generated by the higher-level universal theory (U) grammar. U defines a prior distribution over the space of possible theories, $P(T|U)$, and again the data D determine a likelihood function, with Bayes' rule assigning a posterior probability score to each theory,

$$P(T|D,U) \propto P(D|T)P(T|U). \tag{4.2}$$

Bayes' rule here captures the intuition of Occam's razor, that the theory which best explains a data set (has highest posterior probability $P(T|D,U)$) should balance between fitting the data well (as measured by the likelihood $P(D|T)$), and being simple or short to describe in our general language of thought (as measured by the prior $P(T|U)$). Probabilistic inference can operate in parallel across this hierarchical framework, propagating data-driven information upward and theory-based constraints downward to make optimal probabilistic inferences at all levels simulta-

neously.

Below we explain how each of these probability distributions is defined and computed. The first step is to be more precise about how we represent theories, which we have described so far using an informal mix of logic and natural langauge but now formalize using first-order predicate logic.

A language for theories. Following [86] we choose to represent the laws in a theory as Horn clauses, logical expressions of the form $r \leftarrow (f \wedge g \wedge ... \wedge s \wedge t)$, where each term $r, f, g, s, t, ...$ is a predicate expressing an attribute or relation on entities in the domain, such as $f(X)$ or $s(X, Y)$. Horn clauses express logical implications – a set of conjunctive conditions under which r holds – but can also capture intuitive causal relations [89] under the assumption that any propositions not generated by the theory are assumed to be false. The use of implicational clauses as a language for causal theories was explored extensively in [37].

While richer logical forms are possible, Horn clauses provide a convenient and tractable substrate for exploring the ideas of stochastic search over a space of theories. In our formulation, the Horn clauses contain two kinds of predicates: "core" and "surface". Core predicates are those that cannot be reduced further using the theory's laws. Surface predicates are derived from other predicates, either surface or core, via the laws. Predicates may or may not be directly observable in the data. The core predicates can be seen as compressing the full model into just the minimal bits necessary to specify all true facts. As we explain in more detail below, a good theory is one that compresses a domain well, that explains as much of the observed data as possible using only the information specified in the core predicates. In our magnetism example, the core could be expressed in terms of two predicates $f(X)$ and $g(X)$. Based on an assignment of truth values to these core predicates, the learner can use the theory's laws such as $interacts(X, Y) \leftarrow f(X) \wedge g(Y)$ to derive

values for the observable surface predicate $interacts(X, Y)$. For n objects, there are $O(n^2)$ interactions that can be observed (between all pairs of objects) but these can be explained and predicted by specifying only $O(n)$ core predicate values (for each object, whether or not it is a magnet or is magnetic).

G: Who is your dad?
T: Grandfather Shimon.
G: Oh! Is he my dad too?
T: No.
G: Is Grandma Chana also my grandpa?
T: No, Grandma is your grandma.
G: Ok.
T: And Grandma is my mom.
G: Yeah...that works.
(Conversation with my three-year old)

As another example of how a theory supports compression via its core predicates and abstract laws, consider the domain of kinship as shown in Fig. 4-8. A child learning this domain might capture it by core predicates such as *parent*, *spouse*, and *gender*, and laws such as "Each child has two parents of opposite gender, and those parents are each others' spouse"; "A male parent is a *father*"; "Two individuals with the same parent are *siblings*"; "A female sibling is a *sister*"; and so on. Systems in this domain would correspond to individual families that the child knows about. A system could then be compressed by specifying only the values of the core predicates, for example which members of a family are spouses, who is the parent of whom, and who is male or female. From this minimal set of facts and concepts all other true facts about a particular family can be derived, predicting new relationships that were not directly observed.

In constructing a theory, the learner introduces abstract predicates via new laws, or new roles in existing laws, and thereby essentially creates new concepts. Notice that the core predicates in our magnetism theory need be represented only in purely abstract terms, $f(X)$ and $g(X)$, and initially they have only this bare abstract meaning. They acquire their meaning as concepts picking out magnets or magnetic objects respectively in virtue of the role they play in the theory's laws and the explanations they come to support for the observed data. This is the sense in which our framework allows the introduction of genuinely new abstract concepts via their inferential or conceptual roles.

Entities may be typed and predicates restricted based on type constraints. For

example, in the taxonomy theory shown in Fig. 4-8, $has\_a(X, Y)$ requires that X be a category and Y be a property, while $is\_a(X, Y)$ requires that X and Y both be categories. Forcing candidate models and theories to respect these type constraints provides the learner with another valuable and cognitively natural inductive bias.

Although our focus here is on the acquisition of intuitive theories in general, across all domains of knowledge and all ages, much research has been concerned with the form of young children's theories in a few core domains and the development of that knowledge over the first years of life. Our horn-clause language is too limited to express the full richness of a two-year-old's intuitive physics or intuitive psychology, but it can represent simplified versions of them. For example, in Fig. 4-8 we show a fragment of a simple "desire psychology" theory, one hypothesized early stage in the development of intuitive psychology around two years of age [192]. This theory aims to explain agents' goal-directed actions, such as reaching for, moving towards or looking for various object, in terms of basic but unobservable desires. In our language *desires(X, Y)* (or simply $d(X, Y)$ in Fig. 4-8) is a core predicate relating an agent X to an object Y. Desires are posited to explain observations of a surface predicate $goes\_to(X, Z, S)$: agent X goes to (or reaches for or looks in) location Z in situation S. We also introduce background information in the form of an additional predicate *location(Y,Z,S)* available to the child, specifying that object Y is in location Z in situation S. Then by positing which agents desire which objects, and a law that says effectively, "an agent will go to a certain location in a given situation if that location contains an object that the agent desires", a child can predict how agents will act in various situations, and explain why they do so.

The theory prior $P(T|U)$. We posit U knowledge in the form of a probabilistic

Top level theory

(S1)	S	$\Rightarrow$	(Law) $\wedge$ S
(S2)	S	$\Rightarrow$	(Tem) $\wedge$ S
(S3)	S	$\Rightarrow$	Stop

Random law generation

(Law)	Law	$\Rightarrow$	($F_{left} \leftarrow F_{right} \wedge$ Add)
(Add1)	A	$\Rightarrow$	F $\wedge$ Add
(Add2)	A	$\Rightarrow$	Stop

Predicate generation

($F_{left}1$)	F_{left}	$\Rightarrow$	$surface1()$
$\vdots$			
($F_{left}\ \alpha$)	F_{left}	$\Rightarrow$	$surface\alpha()$
($F_{right}1$)	F_{right}	$\Rightarrow$	$surface1()$
$\vdots$			
($F_{right}\ \alpha$)	F_{right}	$\Rightarrow$	$surface\alpha()$
($F_{right}(\alpha+1)$)	F_{right}	$\Rightarrow$	$core1()$
$\vdots$			
($F_{right}(\alpha+\beta)$)	F_{right}	$\Rightarrow$	$core\beta()$

Law templates

(Tem1)	Tem	$\Rightarrow$	template1()
$\vdots$			
(Temγ)	Tem	$\Rightarrow$	templateγ()

Figure 4-4: Production rules of the Probabilistic Horn Clause Grammar. S is the start symbol and *Law*, *Add*, F and *Tem* are non-terminals. α, β, and γ are the numbers of surface predicates, core predicates, and law templates, respectively.

context-free Horn clause grammar (PHCG) that generates the hypothesis space of possible Horn-clause theories, and a prior $P(T|U)$ over this space (Fig. 4-4). This grammar and the Monte Carlo algorithms we use to sample or search over the theory posterior $P(T|D,U)$ are based heavily on [61], which introduced the approach for learning single rule-based concepts rather than the larger law-based theory structures

we consider here. We refer readers to [61] for many technical details. Given a set of possible predicates in the domain, the PHCG draws laws from a random construction process (Law) or from law templates (Tem; explained in detail below) until the Stop symbol is reached, and then grounds out these laws as horn clauses. The prior $P(T|U)$ is the product of the probabilities of choices made at each point in this derivation. Because all these probabilities are less than one, the prior favors simpler theories with shorter derivations. The precise probabilities of different laws in the grammar are treated as latent variables and integrated out, which favors re-use of the same predicates and law components within a theory [61].

Law templates. We make the grammar more likely to generate useful laws by equipping it with templates, or canonical forms of laws that capture structure likely to be shared across many domains. While it is possible for the PHCG to reach each of these law forms without the use of templates, their inclusion allows the most useful laws to be invented more readily. They can also serve as the basis for transfer learning across domains. For instance, instead of having to re-invent transitivity anew in every domain with some specific transitive predicates, a learner could recognize that the same transitivity template applies in several domains. It may be costly to invent transitivity for the first time, but once found – and appreciated! – its abstract form can be readily re-used. The specific law templates used are described in Fig. 4-5. Each "$F(\cdot)$" symbol stands for a non-terminal representing a predicate of a certain -arity. This non-terminal is later instantiated by a specific predicate. For example, the template $F(X,Y) \leftarrow F(X,Z) \wedge F(Z,Y)$ might be instantiated as $is\_a(X,Y) \leftarrow is\_a(X,Z) \wedge is\_a(Z,Y)$ (a familiar transitive law) or as $has\_a(X,Y) \leftarrow is\_a(X,Z) \wedge has\_a(Z,Y)$ (the other key law of taxonomy, which is like saying that "has_a is transitive over is_a"). This template could be instantiated differently in other domains, for example in kinship as

$child(X, Y) \leftarrow child(X, Z) \wedge spouse(Z, Y)$, which states that the child-parent relationship is transitive over *spouse*.

F(X,Y) ← F(X,Z)∧F(Z,Y)	F(X,Y) ← F(X)∧F(Y)
F(X,Y) ← F(Z,X)∧F(Z,Y)	F(X,Y) ← F(Y,X)
F(X,Y) ← F(X,Z)∧F(Y,Z)	F(X,Y) ← F(X,Y)
F(X,Y) ← F(Z,X)∧F(Y,Z)	F(X) ← F(X)
F(X,Y) ← F(X,Y)∧F(X)	F(X) ← F(X,Y)∧F(X)
F(X,Y) ← F(Y,X)∧F(X)	F(X) ← F(Y,X)∧F(X)
F(X,Y) ← F(X,Y)∧F(Y)	F(X) ← F(X,Y)∧F(Y)
F(X,Y) ← F(Y,X)∧F(Y)	F(X) ← F(Y,X)∧F(Y)

Figure 4-5: Possible templates for new laws introduced by the grammar. The leftmost F can be any surface predicate, the right F can be filled in by any surface or core predicates, and X and Y follow the type constraints.

The theory likelihood $P(D|T)$. An abstract theory makes predictions about the observed data in a domain only indirectly, via the models it generates. A theory typically generates many possible models: even if a child has the correct theory and abstract concepts of magnetism, she could categorize a specific set of metal bars in many different ways, each of which would predict different interactions that could be observed as data. Expanding the theory likelihood,

$$P(D|T) = \sum_M P(D|M)P(M|T), \tag{4.3}$$

we see that theory T predicts data D well if it assigns high prior $P(M|T)$ to models M that make the data probable under the observation process $P(D|M)$.

The model prior $P(M|T)$ reflects the intuition that a theory T explains some data well if it compresses well: if it requires few additional degrees of freedom beyond its abstract concepts and laws – that is, few specific and contingent facts about the system under observation, besides the theory's general prescriptions – to make its predictions. This intuition is captured by a prior that encourages the core predicates to be as sparse as possible, thereby penalizing theories that can only fit well by "overfitting" with many extra degrees of freedom. This sparseness assumption is reasonable as a starting point for many domains, given that core predicates are meant to explain and compress the data. Formally, we assume a conjugate beta prior on all binary facts in M, modeled as Bernoulli random variables which we integrate out analytically, as in [86].

Finally, the model likelihood $P(D|M,T)$ comes from assuming that we are observing randomly sampled true facts (sampled with replacement, so the same fact could be observed on multiple occasions), which also encourages the model extension to be as small as possible. This provides a form of implicit negative evidence [175], useful as an inductive bias when only positive facts of a domain are observed.

Stochastic search in theory space: a grammar-based Monte Carlo algorithm. Following [61], we use a grammar-based Metropolis-Hastings (MH) algorithm to sample theories from the posterior distribution over theories conditioned on data, $P(T|D,U)$. This algorithm is applicable to any grammatically structured theory space, such as the one generated by our PHCG; it is also a version of the Church MH inference algorithm [58]. The MH algorithm is essentially a Markov chain on the space of potential derivations from the grammar, where each step in the chain – each proposed change to the current theory – corresponds to regenerating some sub-

tree of the derivation tree from the PHCG. For example, if our theory of magnetism includes the law $interacts(X, Y) \leftarrow f(X) \wedge g(Y)$, the MH procedure might propose to add or delete a predicate (e.g., $interacts(X, Y) \leftarrow f(X) \wedge g(Y) \wedge h(Y)$ or $interacts(X, Y) \leftarrow f(X)$), to change one predicate to an alternative of the same form (e.g., $interacts(X, Y) \leftarrow f(X) \wedge h(Y)$) or a different form if available (e.g., $interacts(X, Y) \leftarrow f(X) \wedge s(X, Y)$); to resample the law from a template (e.g., $interacts(X, Y) \leftarrow t(X, Z) \wedge t(Z, Y)$); or to add or delete a whole law.

These proposals are accepted with probability equal to the maximum of 1 and the MH acceptance ratio,

$$\frac{P(T'|D, U)}{P(T|D, U)} \cdot \frac{Q(T|T')}{Q(T'|T)}, \tag{4.4}$$

where T is the current theory, T' is the new proposed theory, and $Q(\cdot|\cdot)$ is the transition probability from one theory to the other, derived from the PHCG [61]. To aid convergence we raise these acceptance ratios to a power greater than 1, which we increase very slightly after each MH step in a form of simulated annealing. Early on in learning, a learner is thus more likely to try out a new theory that appears worse than the current one, exploring candidate theories relatively freely. However, with time the learner becomes more conservative – increasingly likely to reject new theories unless they lead to an improved posterior probability.

While this MH algorithm could be viewed merely as a way to approximate the calculations necessary for a hierarchical Bayesian analysis, we suggest that it could also capture in a schematic form the dynamic processes of theory acquisition and change in young children. Stochastic proposals to add a new law or change a predicate within an existing law are consistent with some previous characterizations of children's theory learning dynamics [158]. These dynamics were previously proposed

on purely descriptive grounds, but here they emerge as a consequence of a rational learning algorithm. Although the dynamics of an MH search might appear too random to an omniscient observer who knows the "true" target of learning, it would not be fair to call the algorithm sub-optimal, because it is the only known general-purpose approach for effectively searching a complex space of logical theories. Likewise, the annealing process that leads learning to look child-like in a certain sense – starting off with more variable, rapidly changing and adventurous theories, then becoming more conservative and less variable over time – also makes very good engineering sense. Annealing has proven to be useful in stochastic search problems across many scientific domains [97] and is the only known method to ensure that a stochastic search converges to the globally optimal solution. It does not seem implausible that some cognitive analog of annealing could be at work in children's learning.[3]

Approximating the theory score: an inner loop of MCMC Computing the theory likelihood $P(D|T)$, necessary to compare alternative theories in Equation (4.4), requires a summation over all possible models consistent with the current theory (Equation (4.3)). Because this sum is typically very hard to evaluate exactly, we approximate $P(D|T)$ with $P(D|M^*, T)P(M^*|T)$, where M^* is an estimate of the maximum a-posteriori (MAP) model inferred from the data: the most likely values of the core predicates. The MAP estimate M* is obtained by running an inner sampling procedure over the values of the core predicates. As in [86], we use a specialized form of Metropolis-Hastings sampling known as Gibbs sampling. The Gibbs sampler goes over each core predicate assignment in turn while keeping all other assignments

[3]It is worth noting that annealing could be implemented in a learning system without an explicit temperature parameter or cooling schedule, merely based on experience accumulating over time. Here for simplicity we have kept the learner's dataset fixed, but if the learner is exposed to increasing amounts of data over time and treats all data as independent samples from the model, this also acts to lower the effective temperature by creating larger ratios between likelihoods (and hence posterior probabilities) for a given pair of theories.

fixed, and proposes changes to the currently considered assignment. As a concrete example of how the Gibbs loop works, consider a learner who is proposing a theory that contains the law $interacts(X, Y) \leftarrow f(X) \wedge g(Y)$, i.e., objects for which core predicate f is true interact with objects for which core predicate g is true. The learner begins by randomly extending the core categories over the domain's objects: e.g., f might be posited to hold for objects 1, 4, and 7, while g holds for objects 2, 4, 6, and 8. (Note how either, both or none of the predicates may hold for any object, a priori.) The learner then considers the extension of predicate f and proposes removing object 1, scoring the new model (with all other assignments as before) on the observed data and accepting the proposed change probabilistically depending on the relative scores. The learner then considers objects 2, 3, and so on in turn, considering for each object whether predicate f should apply, before moving on to predicate g. (These object-predicate pairs are often best considered in random order on each sweep through the domain.) This process continues until a convergence criteria is reached. We anneal slightly on each Gibbs sweep to speed convergence and lock in the best solution. The Gibbs sampler over models generated by a given theory is thus an "inner loop" of sampling in our learning algorithm, operating within each step of an "outer loop" sampling at a higher level of abstract knowledge, the MH sampler over theories generated by U knowledge.

4.4 Case Studies

We now explore the performance of this stochastic approach to theory learning in two case studies, using simulated data from the domains of taxonomy and magnetism introduced above. We examine the learning dynamics in each domain and make more explicit the possible parallels with human theory acquisition.

4.4.1 Taxonomy

As we saw earlier, the domain of taxonomy illustrates how a compressive knowledge representation is useful in capturing semantic data. How can such a powerful organizing principle itself be learned? [86] showed that a Bayesian ideal observer can pick out the best theory of taxonomy given a small set of eight possible alternatives. Here we show that the theory of taxonomy can be learned in a more constructive way, via an MCMC search through our infinite grammar-generated hypothesis space. The theory to be learned takes the following form:

Two core predicates:	$s(X,Y)$ and $t(X,Y)$
Two observable predicates:	$is\_a(X,Y)$ and $has\_a(X,Y)$

Law 1:	$is\_a(X,Y) \leftarrow s(X,Y)$
Law 2:	$has\_a(X,Y) \leftarrow t(X,Y)$
Law 3:	$is\_a(X,Y) \leftarrow is\_a(X,Z) \wedge is\_a(Z,Y)$
Law 4:	$has\_a(X,Y) \leftarrow is\_a(X,Z) \wedge has\_a(Z,Y)$

These laws by themselves do not yet capture the complete knowledge representation we are after; we also need to instantiate the core predicates in a particular model. These laws allow many possible models for any given data sets. One of these models is the compressed tree representation (shown in Fig. 4-8 in the Model section of the taxonomy domain), which specifies only the minimal facts needed to derive the observed data from Laws 1-4. A different model could link explicitly all the *is_a(X, Y)* connections, for example drawing the links between *salmon* and *animal*,

shark and *animal* and so on. Another model could link explicitly all the *has_a(X, Y)* connections. However, these latter two models would be much less sparse than the compressed tree representation, and thus would be disfavored relative to the compressed tree shown in Fig. 4-8, given how we have defined the model prior $P(M|T)$. In sum, in this framework, the organization of categories and properties into a tree-structured inheritance hierarchy comes about from a combination of positing the appropriate abstract laws and core predicates together with a sparsity preference on the assignments of the core predicates' values.

Note also that the core predicates $s(X, Y)$ and $t(X, Y)$ acquire their meaning in part by their inferred extensions, and in part by how they are related to the observed surface predicates. The surface predicates are assumed to be verbal labels which the learner observes and needs to account for. The link between these verbal labels and the core relations are what given by Laws 1 and 2. While these links could in general also be learned, we follow [86] in taking Laws 1 and 2 as given for this particular domain and asking whether a learner can theoretically discover Laws 3 and 4 – but now at the algorithmic level. We test learning for the same simple model of the taxonomy domain studied by Katz et al., using seven categories and seven properties in a balanced tree structure. We presented all true facts from this model as observations to the learner, including both property statements (e.g., "An eagle has claws") and category membership statements (e.g., "An eagle is a bird"). The data for this section and the following case study can be found in the appendix.

We ran 60 simulations, each comprising 1300 iterations of the outer MH loop (i.e., moves in the space of theories). Four representative runs are shown in Fig. 4-6, as well as the average across all the runs. Out of 60 simulations, 52 found the correct theory within the given number of iterations, and 8 discovered a partial theory which included only Law 3 or Law 4.

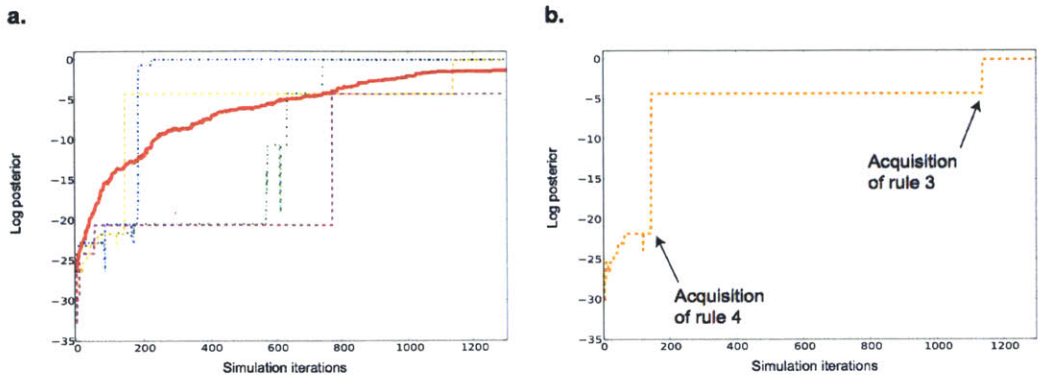


Figure 4-6: Representative runs of theory learning in Taxonomy. **(a)** Dashed lines show different runs. Solid line is the average across all runs. **(b)** Highlighting a particular run, showing the acquisition of law 4, followed by the acquisition of law 3 and thus achieving the final correct theory.

Several points are worth noting beyond these quantitative results. First, it is striking that abstract structure can be learned effectively from very little data. Using simple local search, our learning algorithm is able to navigate an infinite space of potential theories and discover the true laws underlying the domain, even with relatively few observations in the relations between seven categories and seven properties. This is a version of the "blessing of abstraction" described in [63] and [178], but one that is realized at the algorithmic level and not just the computational level of ideal learning.

Second, individual learning trajectories proceed in a characteristic pattern of stochastic leaps. Discovering the right laws gives the learner strong explanatory power. However, surrounding each "good" theory in the discrete hypothesis space are many syntactically similar but nonsensical or much less useless formulations. Moving from a good theory to a better one thus depends on proposing just the right changes to the current hypothesis. Since these changes are proposed randomly, the learner

often stays with a particular theory for many iterations, rejecting many proposed alternatives which score worse or not significantly better than the current theory, until a new theory is proposed that is so much better it is almost surely accepted. This leads to the observed pattern of plateaus in the theory score, punctuated by sudden jumps upward and occasional jumps downward in probability. While we do not want to suggest that people learn theories only by making random changes to their mental structures, the probabilistic nature of proposals in a stochastic search algorithm could in part explain why individual human learning curves rarely proceed along a smooth path and can show broad variation across individuals given the same data.

Third, while individual learning trajectories may be discontinuous, on average learning appears smooth. Aggregating performance over all runs shows a smooth improvement of the theory's score that belies the underlying discrete nature of learning at an individual level. This emphasizes the possible danger of studying theory learning and theory change only in the average behavior of groups of subjects, and the theoretical value of microgenetic methods [159] for constraining algorithmic-level models of children's' learning.

4.4.2 Magnetism

We now turn to the domain of magnetism, where the trajectory of theory learning reveals not only successful acquisition, but interesting intermediate stages and transitions corresponding to classic phenomena of conceptual change in childhood and early science [26]. The simplified theory of magnetism to be learned takes the following form:

Two core predicates:	$f(X)$ and $g(X)$
One observable predicate:	$interacts(X, Y)$

Law 1:	$interacts(X, Y) \leftarrow f(X) \wedge f(Y)$
Law 2:	$interacts(X, Y) \leftarrow f(X) \wedge g(Y)$
Law 3:	$interacts(X, Y) \leftarrow interacts(Y, X)$

The particular model used for learning contained 10 objects: 3 magnets, 5 magnetic objects and 2 non-magnetic objects. The learner was given all true facts in this model, observing interactions between each magnet and every other object that was either a magnet or a magnetic object, but no other interactions. Unlike in the previous taxonomy example, the learner was given none of the laws or core predicate structure to begin with; the entire theory had to be constructed by the search algorithm. Assuming the correct laws (as shown above) can be found, the model prior $P(M|T)$ favoring sparsity suggests the optimal values for the core predicates should assign one core predicate (f) to all and only the magnets, and another predicate (g) to all and only the non-magnet magnetic objects. This leads to the theory and model depicted jointly in Fig. 4-8.

We ran 70 simulations, each comprising 1600 iterations of the outer MH loop sampling over candidate theories. In many respects, results in this domain were qualitatively similar to what we described above for taxonomy. Out of 70 simulated learning runs, 50 found the correct theory or a minor logical variant of it; the rest discovered a partial theory. The correct final theories account for the full observed data and only the observed data, using three laws. While all the full theories learned included Laws 1 and 2, only some of them included the exact form of Law 3, express-

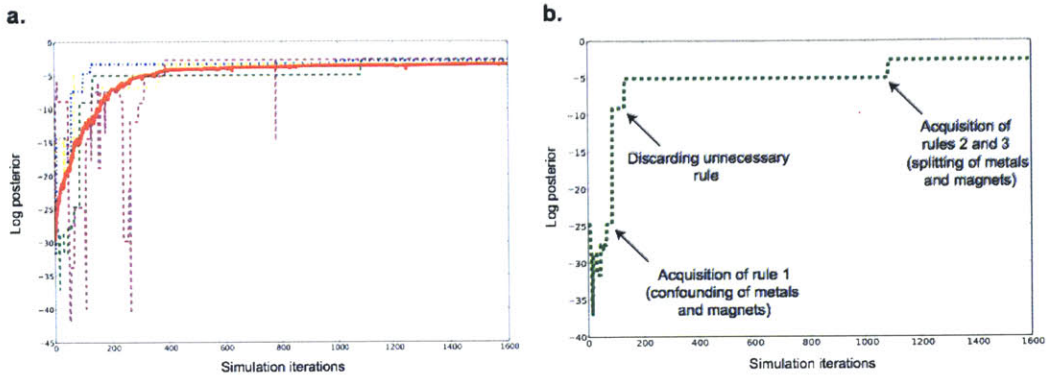


Figure 4-7: Representative runs of theory learning in Magnetism. **(a)** Dashed lines show different runs. Solid line is the average across all runs. **(b)** Highlighting a particular run, showing the acquisition of law 1 and the confounding of magnets and magnetic (but non-magnet) objects, the discarding of an unnecessary law which improves the theory prior, and the acquisition of the final correct theory.

ing the symmetry of interaction.[4] The dynamics of representative runs are displayed in Fig. 4-7, as well as the average over all the runs. As in the domain of taxonomy, individual learners experienced radical jumps in their theories, while aggregating across runs learning appears to be much smoother.

The most interesting aspects of learning here were found in the transitions between distinct stages of learning, when novel core predicates are introduced and existing core predicates shift their meaning in response. Key transitions in children's cognitive development may be marked by restructuring of concepts, as when one core concept differentiates into two [23]. Our learning algorithm often shows this same dynamic in the magnetism task. There is no single order of concept acquisition that

[4]However, the variants discovered were functionally equivalent within this domain to symmetry. Such variants include redundant re-statements of symmetry, such as interacts(X,Y) $\leftarrow$ interacts(Y,Z) $\wedge$ equals(Z,X). Other forms happen to capture the same facts as symmetry within this particular domain, such as interacts(X,Y) $\leftarrow$ interacts(Y,Z) $\wedge$ g(Z). These variants appear more complex than the basic symmetry law, and they do score slightly worse than theories that recover the original formulation. However, since they were generated by templates in this case, this extra complexity does not hurt them significantly.

the algorithm follows in all or most runs, but the most common trajectory (shown in Fig. 4-7b) involves learning Law 1 first, followed later by the acquisition of Laws 2 and 3. As mentioned earlier, for a learner who knows only Law 1, the optimal setting of the core predicates is to lump together magnets and magnetic objects in one core predicate, essentially not differentiating between them. Only when Laws 2 and 3 are learned does the learner also acquire a second core predicate that carves off the magnetic non-magnets from the magnets. On a smaller number of runs, a different order of acquisition is observed: first Laws 2 and 3 are learned, and then Law 1 is added. This sequence also involves a conceptual restructuring, albeit a less dramatic one. A learner who possesses only Laws 2 and 3 will optimally assign one predicate to all and only the magnets, and another core predicate to both magnets and magnetic non-magnets, again lumping these two classes together. Only once Law 1 is added to Laws 2 and 3 will the learner completely differentiate the two core predicates with non-overlapping extensions corresponding to magnets and magnetic non-magnets.

In both of these cases, the time course of learning appears as a progression from simpler theories (with fewer core predicates and/or laws) that explain the data less faithfully or less efficiently, to more complex theories (with more core predicates and/or laws) that explain the data more faithfully or more efficiently. A learner with the simpler theory consisting of only Law 1 (without Laws 2 and 3) will overgeneralize, predicting the existence of interactions that do not actually occur: interactions between pairs of non-magnet magnetic objects (which would be treated the same as interactions between two magnets, or a magnet and a magnetic object). A learner with the simpler theory consisting of Laws 2 and 3 (but not Law 1) will make the right predictions about interactions to be observed but would represent the world less efficiently, less sparsely, than they could: they would need to assign values for

both core predicates to represent each magnet, rather than just using a single core predicate to represent magnets and only magnets. Yet while being less accurate or less efficient, these earlier, simpler theories are still reasonable first approximations to the optimal theory of this domain. They are also plausible intermediate points for the learner on the way to the optimal theory, who can get there merely by adding one or two new laws and differentiating the extension of a core predicate into two non-overlapping subsets of objects, magnets and magnetic non-magnets, which had previously been merged together in that predicate's extension.

4.5 Two Sources of Learning Dynamics

"Had we but world enough, and time" – Andrew Marvell, To His Coy Mistress

The story of development is in essence one of time and data. In order to construct adult-level intuitive theories, children require both sufficient time to ponder and exposure to sufficient evidence. For a child on the verge of grasping a new theory, either additional data or additional time to think can make the difference [26]. Measured as a function of either time or amount of data experienced, the dynamics of learning typically follows an arc from simpler theories that only coarsely predict or encode experience to more complex theories that more faithfully predict and encode it. The above case studies of theory learning in the domains of taxonomy and magnetism show this dynamic as a function of time elapsed in the search process, for a fixed data set. Previous Bayesian models of theory learning [91] have emphasized the complementary perspective: how increasing amounts of data naturally drive an ideal learner to more complex but more predictive theories, independent of the dynamics of search or inference.

These two sources of learning dynamics are most naturally understood at different levels of analysis. Data-driven learning dynamics seems best explained at the

computational level, where the ideal learner shifts probability mass between candidate theories as a function of the data observed. In contrast, time-driven dynamics (independent of the amount of data observed) seems best approached at the algorithmic level, with models that emphasize how the learner's process of searching over a hypothesis space unfolds over time independent of the pace with which data accumulates.

Our modeling approach is well suited to studying both data-driven and time-driven dynamics and their interactions, because of its focus on the interface between the computational and algorithmic levels of analysis. In the rest of this section we return to the domain of simplified magnetism and explore the independent effects and interactions of these two different drivers of theory change in our model. How does varying time and data affect our ideal learner? We provide the learner with several different data sets, and examine how the learning dynamics unfold over time for each one of these sets. In each data set we provide the learner with different observations by parametrically varying the number of magnetic objects over five cases, which can be ordered in the following way: Case 1 had 3 magnets, 1 magnetic object and 6 non-magnetic objects. Each case then adds one magnetic object while removing one non-magnetic object, so that case 2 has 3 magnets, 2 magnetic objects and 5 non-magnetic objects, up to case 5 which has 3 magnets, 5 magnetic objects and 2 non-magnetic objects (the same as the previous section). We also considered a special case, case X, in which there is only 1 magnet, 7 magnetic objects and 2 non-magnetic object. In all cases the theory governing the domain is exactly the same as that described in the magnetism case study. Given these different cases we find that at the end of the simulation the learner almost always settled on one of three theories. We therefore focus on these three theories, the formal laws of which are given in Fig. 4-8a. Informally, these theories correspond to:

Theory A: "There is one class of interacting objects in the world, and objects in this class interact with other objects in this class."

Theory B: "There are two classes of interacting objects in the world, and objects from one class interact with objects in the other class. These interactions are symmetric."

Theory C: "There are two classes of interacting objects in the world, and objects from one class interact with objects in the other class. Also, objects in one of the classes interact with other objects in the same class. These interactions are symmetric."

It is important to emphasize that theories A, B and C were not given to the learner as some sort of limited hypothesis space. Rather, the number of possible theories the learner could consider in each case is potentially infinite, but practically it settles on one of these three or their logical equivalents. Many other theories besides A, B and C were considered by the learner, but they do not figure significantly into the trajectory of learning. These theories are much less good (i.e., unnecessarily complex or poorly fitting) relative to neighboring knowledge states, so they tend to be proposed and accepted only in the early, more random stages of learning, and are quickly discarded. We could not find a way to group these other theories into cohesive or sensibly interpreted classes, and since they are only transient states of the learner, we removed them for purposes of analyzing learning curves and studied only the remaining proportions, renormalized.

In order to see how the dynamics of learning depend on data, consider specifically cases in which there are few magnetic objects that are not magnets, perhaps 1 or 2 (as in cases 1 and 2). In this case a partial theory such as theory A might suffice. According to this theory there is only one type of interacting object, and one law. If there are two magnetic non-magnets in the domain, the partial theory will classify

them as 'interacting' objects based on their behavior with the magnets, conflating them with the magnets. However, it will incorrectly predict the two magnetic non-magnets should interact with each other. Their failure to interact will be treated as an outlier by the learner who has theory A. The full theory C can correctly predict this non-interaction, but it does so by positing more laws and types of objects, which has a lower prior probability. As the number of magnetic non-magnets increases, the number of 'outliers' in the data increase as well (see Fig. 4-8b). Theory A now predicts more and more incorrect interactions, and in a Kuhnian fashion there is a point at which these failures can no longer be ignored, and a qualitative shift to a new theory is preferred. In a completely different scenario, such as the extreme case of only 1 magnet (case X), we might expect the learner to not come up with magnet interactions laws, and settle instead on theory B.

For each one of the outlined cases we ran 70 simulations for 1600 iterations. Fig. 4-8c shows the effect of data and time on the learning process, by displaying the relative proportion of the outlined theories at the end of the iteration for all simulations. Note the transition from case 1 to case 5: With a small number of non-magnet magnetic objects, the most frequently represented theory is theory A, which puts all magnetic objects (magnet or not) into a single class and treats the lack of interactions between two magnetic non-magnets as essentially noise. As the number of magnetic non-magnets increases, the lack of interactions between the different non-magnets can no longer be ignored and the full theory becomes more represented. Case X presents a special scenario in which there is only 1 magnet, and as expected theory B is the most represented there. The source of the difference between the proportion of theories learned in these different cases is the data the learner was exposed to. Within each case, the learner undergoes a process of learning similar to that described in the case studies – adopting and discarding theories in a process or

time-driven manner.

To summarize, theory acquisition can be both data-driven and process-driven. Our simulations suggest that, at least in this simplified domain, both sufficient data and sufficient time to think are required. Only when the observed data provide a strong enough signal – as measured here by potential outliers under a simpler theory – is there sufficient inductive pressure for a Bayesian learner guided by simplicity priors to posit a more complex theory. Yet even with all the data in the world, a practical learning algorithm still requires sufficient time to think, time to search through a challenging combinatorial space of candidate laws and novel concepts and construct a sequence of progressively higher scoring theories that will reliably converge on the highest scoring theory for the domain.[5] The fact that both sufficient data and sufficient time are needed for proper theory learning fits with the potentially frustrating experience of many teachers and parents: having laid out for a child all the data, all the input, that they need to solve a problem, grasp some explanation, or make a discovery, the child still doesn't seem to get it, or takes surprisingly long to get it. Knowing that any realistic learner needs both data enough, and time, may at least provide some relief from that frustration and the patience to watch and wait as learning does its work.

4.6 Evidence from experiments with children

While our work here has been primarily motivated by theoretical concerns, we also want to consider the empirical evidence that children's learning corresponds in some way to the computational picture we have developed. Our most basic result is that a

[5]It should however be noted that in some cases, the time-component allows the learner to 'weed out' and abandon overly complex theories.

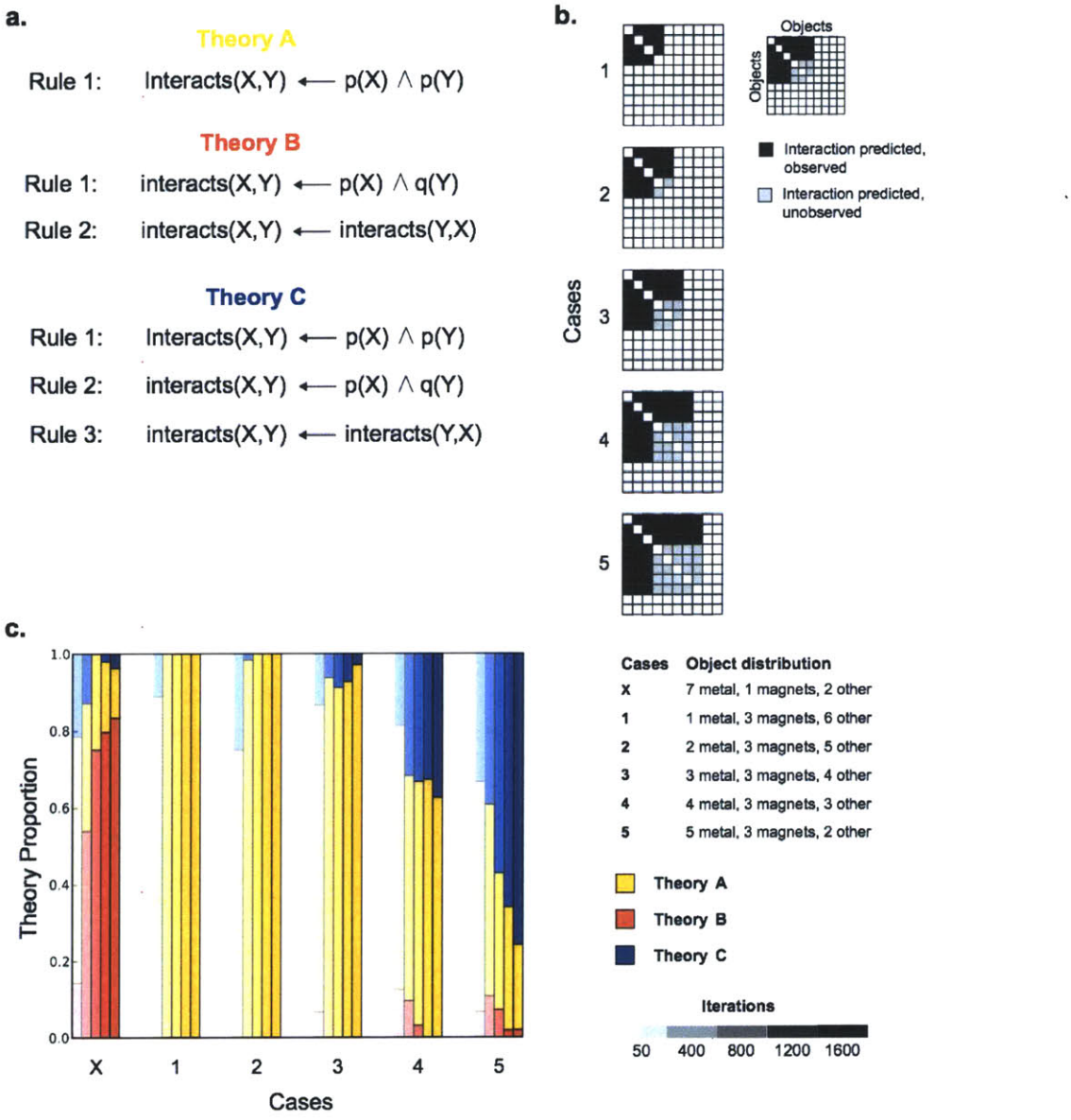


Figure 4-8: Learning dynamics resulting from two different sources: **(a)** A formal description of theories A, B and C **(b)** The predicted and observed interactions given theory A for the different cases, showing the growing number of outliers as the number of magnetic non-magnet objects grows **(c)** Proportion of theories accepted by the learner for different cases, during different points in the simulation runs. More opaque bars correspond to later iterations in the simulation. Different theories are acquired as a result of varying time and data.

simple, cognitively plausible, stochastic search algorithm, guided by an appropriate grammar and language for theories, is capable of solving the rather sophisticated joint inference problem of learning both the concepts and the laws of a new theory – what we referred to as the "hard problem" or the "chicken-and-egg" problem of theory learning. In the last few years, several lines of experimental work have shown that children and adults can indeed solve this joint inference problem in the course of acquiring new theories. [92] showed that adults were able to learn new causal relations, such as *objects of type A light up type B*, and to use these relations to categorize objects, for example *object 3 is of type A*. In [107] adults performed a task asking about specific causal structures leading to evidence (which objects are 'blickets' that cause a 'blicket-meter' to activate), which required inferring the abstract functional form of the causal relations (do blickets activate the meter via a noisy-OR function, a deterministic disjunctive function or a conjunctive function). A similar experiment [108] demonstrated that children are also able to acquire such abstract knowledge about the functional causal form while considering the specific identity of objects.

While in these studies children were explicitly told that only one type of concept is involved, Schulz and colleagues [154] showed that young children can solve an even more challenging task: Given sparse evidence in the form of different blocks touching and making different noises, the children correctly posited the existence of three different causal kinds underlying the observed relations. In this case the children had to both infer the abstract relations governing the behavior, and posit how many concepts underly these relations. These papers are qualitatively consistent with the predictions of our approach . In [17] we showed a more quantitative correspondence between our model predictions and children's categorization judgments. In that study children were shown interactions in a domain of simplified magnetism, where

several unlabeled blocks interacted with blue and yellow blocks, either attracting or repelling from them. We also showed that the Monte Carlo search algorithm given here is capable of finding just the theories that children do, or theories that are behaviorally indistinguishable from them, and revising them appropriately.

Could models from an alternative paradigm such as connectionism also explain these results? Connectionist architectures could potentially solve aspects of the tasks described in [107], for example. There are certainly networks capable of distinguishing between different functional forms like those in [107], which may be seen as learning governing laws in a theory. Connectionist networks can also form new concepts - in the sense of clusters of data that behave similarly - via competitive learning. However, it has yet to be shown that a connectionist network can learn or represent the kinds of abstract knowledge that our approach does, and that children grasp in the other experiments cited above: solving the joint inference problem of discovering a system of new concepts and laws that together explain a set of previously unexpected interactions or relations. This problem poses an intriguing open challenge for connectionist modelers in cognitive development, one that could stimulate significant new research.

Going forward, we would like more fine-grained tests of whether and how the Monte Carlo search learning mechanism we have posited corresponds to the mechanisms by which children explore their space of theories. This will be challenging, as most of the steps of learning are not directly observable. We are currently working on studies together with Bonawitz and colleagues to test some general predictions of our model, such as the trade-off between data and time described in the section above. In these experiments we recreate the domain of simplified magnetism described in the case studies section, with three types of objects that interact according to several laws. The children will be given different amounts of evidence, and crucially different

segments of time, after which they will be asked to sort the objects they see into categories and describe why they do so. The children will not be told in advance how many object types exist, and we anticipate the number of types posited by the children will depend on their current domain theory. We anticipate the same amount of evidence but varying lengths of time will lead children to transition from one theory to the next, which will be evidenced in their sorting behavior. This behavior will be matched with running the stochastic search algorithm for varying amounts of time, as described in the previous section, though we recognize these are still only indirect tests of the model's predictions.

More precision could come from microgenetic methods [159], which study developmental change by giving children the same task several times and inspecting the strategies used to solve the task at many intervals. Microgenetic studies find that often, while the task itself remains constant, the strategies used to solve it undergo change. This data could be interpreted as a search process unfolding over time. A fundamental question for the microgenetic method remains why and how change occurs. Our algorithmic approach offers an explanation of how, and can potentially address the why. Together with Bonawitz and colleagues we are developing micro genetic methods to test whether children's learning can be explained in terms of Monte Carlo search.

One key challenge in designing a microgenetic study is defining an externally measurable sign of the internal cognitive mechanism of hypothesis testing and discovery. Similar to how microgenetic studies keep a task fixed, we intend to observe how children play and experiment with a given set of objects, without introduce new objects or any new data in the form of new interactions that haven't been observed before. As in classic microgenetic studies, we intend to ask the children questions and encourage them to talk out loud about their hypotheses in, order to probe the

state of their search at more abstract levels of the theory. We can score the theories they uncover using computational tools, and observe whether the pattern of theories abandoned, adopted and uncovered fits with Monte Carlo search.

4.7 Discussion and Conclusion

Not all those who wander are lost

– J.R.R Tolkien, All That is Gold Does Not Glitter

We have presented an algorithmic model of theory learning in a hierarchical Bayesian framework and explored its dynamics in several case studies. We find encouraging the successful course of acquisition for several example theories, and the qualitative parallels with phenomena of human theory acquisition. These results suggest that previous "ideal learning" analyses of Bayesian theory acquisition can be approximately realized by a simple stochastic search algorithm and thus are likely well within the cognitive grasp of child learners. It is also encouraging to think that state-of-the-art Monte Carlo methods used in Bayesian statistics and artificial intelligence to approximate ideal solutions to inductive inference problems might also illuminate the way that children learn. At this intersection point between the computational level and the algorithmic level of analysis, we showed that theory change is expected to be both data-driven and process-driven. This is an important theoretical distinction, but the psychological reality of these two sources of learning dynamics and their interaction needs to be further studied in experiments with children and adults.

While the main contributions of this chapter are in addressing the algorithmics of theory acquisition, the 'how', the introduction of law templates provides some

insight regarding 'what' the structure of children's knowledge might be, and the coupling between how we answer 'what?' and 'how?' questions of learning. On an algorithmic level, we found such templates to be crucial in allowing learning to converge on a reasonable timescale. On a computational level, these templates can be seen as generalizing useful abstract knowledge across domains, and providing high-level constraints that apply across all domain theories. The formal framework section did not directly treat where such templates come from, but it is possible to imagine that some of them are built in as overarching constraints on knowledge. More likely, though, they are themselves learned during the algorithmic acquisition process. An algorithmic grammar-based model can learn templates by abstracting successful rules from their particular domain instantiation. That is, if the model (or child) discovers a particularly useful rule involving a specific predicate such as "if *is_a*(X,Y) and *is_a*(Y,Z), then *is_a*(X,Z)", then the specific predicate might be abstracted away to form the transitive template "if F(X,Y) and F(Y,Z), then F(Y,Z)". Learning this transitive template then allows its reuse in subsequent theory, and represents a highly abstract level of knowledge.

There are many ways in which our modeling work here can and should be extended in future studies. The algorithm we have explored is only one particular instance of a more general proposal for how stochastic search operating over a hierarchically structured hypothesis space can account for theory acquisition. The specific theories considered here were only highly simplified versions of the knowledge children have about real-world domains. Part of the reason that actual concepts and theories are richer and more complex is due to the fact that children have a much richer underlying language for representations. Horn clauses are expressive and suitable for capturing some knowledge structures, and in particular certain kinds of causal relations, but they are not enough. A potentially more suitable theory space would

be built on a functional language, in which the laws are more similar to mathematical equations. Such a space would be harder to search through, but it would be much more expressive. A functional language of this sort would allow us to explore rich theories described in children, such as basic notions about objects and their interactions [169], and the intuitive physics of object behavior [5]. Despite the need for a more expressive language, we expect the same basic phenomena found in the model domains considered here to be replicated in more complex models. Moving forward, a broader range of algorithmic approaches, stochastic as well as deterministic, need to be evaluated as both as behavioral models and as effective computational approximations to the theory search problem for larger domains.

Relative to previous Bayesian models of cognitive development that focused on only the computational level of analysis, this chapter has emphasized algorithmic-level implementations of a hierarchical Bayesian computational theory, and the interplay between the computational and algorithmic levels. We have not discussed at all the level of neural implementation, but recent proposals by a number of authors argue that analogous stochastic-sampling ideas could plausibly be used to carry out Bayesian learning in the brain [40]. More generally, a "top-down" path to bridging levels of explanation in the study of mind and brain, starting with higher, more functional levels and moving down to lower, more mechanistic levels, appears most natural for Bayesian or other "reverse-engineering" approaches to cognitive modeling [72]. Other paradigms for cognitive modeling adopt different ways to navigate the same hierarchy. Connectionist approaches, for instance, start from hypothesized constraints on neural representations (e.g., distributed codes) and learning mechanisms (e.g., error-driven learning) and move up from there, to see what higher-level phenomena emerge [115]. While we agree that actual biological mechanisms will ultimately be a central feature of any account of children's cognitive development,

"[T]here seems to be little predictive extrapolation from the 'component' level to the 'computational' level. Extrapolation in the other direction is, however, somewhat easier"
– Marr and Poggio

we are skeptical that this is the best place to start [72]. The details of how the brain might represent or learn knowledge such as the abstract theories we consider here remain largely unknown, making a bottom-up emergent alternative to our approach hard to contemplate. In contrast, while our top-down approach has yet to make contact with neural phenomena, it has yielded real insights spanning levels. Moving from computational-level accounts to algorithms that explicitly (if approximately) implement the computational theory let us see plainly how the basic representations of children's theories could be acquired, and suggest explanations for otherwise puzzling features of the dynamics of learning in young children, as the consequences of efficient and effective algorithms for approximating the rational computational-level ideal of Bayesian learning. We hope that as neuroscience learns more about the neural substrates of symbolic representations and mechanisms of exploratory search, our top-down approach can be meaningfully extended from the algorithmic level to the level of implementation in the brain's hardware.

Going back to the puzzle of the "chicken and egg" problem posed at the beginning of the chapter, what do the dynamics explored here tell us about the coupled challenges of learning the laws of a theory and the invention of truly novel concepts, and the opposing views represented by Fodor and Carey? There is a sense in which, at the computational level, the learner already must begin the learning process with all the laws and concepts needed to represent a theory already accessible. Otherwise the necessary hypothesis spaces and probability distributions for Bayesian learning could not be defined. In this sense, Fodor's skepticism on the prospects for learning or constructing truly novel concepts is justified. Learning cannot really involve the discovery of anything "new", but merely the changing of one's degree of belief in a theory, transporting probability mass from one part of the hypothesis space to another. However, on the algorithmic level explored in this chapter, the level of

active processing for any real-world learner, there is in fact genuine discovery of new concepts and laws. Our learning algorithm can begin with no explicitly represented knowledge in a given domain – no laws, no abstract concepts with any non-trivial extensions in the world – and acquire reasonable theories comprised of novel laws and concepts that are meaningfully grounded and predictively useful in that domain.

Our specific algorithm suggests the following account of how new concepts derive their meanings. Initially, the concepts themselves are only blank predicates. The theory prior induces a non-arbitrary structure on the space of possible laws relating these predicates, and in that sense can be said to contain a space of proto-meanings. The data are then fused with this structure in the prior to create a structured posterior: the concepts are naturally extended over the observed objects in those regions where the posterior has a high probability, and those are the areas in theory space that the learner will converge towards. This algorithmic process is, we suggest, an instance (albeit a very simple one) of Carey's "bootstrapping" account [25, 26] of conceptual change, and a concrete computational implementation of concept learning under an inferential role semantics.

Under Carey's account of the origins of new concepts, children first use symbols as placeholders for new concepts and learn the relations between them that will support later inferential roles. Richer meaning is then filled in on top of these placeholders and relations, using a "modeling process" involving a range of inductive inferences. The outer loop of our algorithm explains the first stage: why some symbolic structures are used rather than others and how their relations are created. The second stage of Carey's account parallels the inner loop of our algorithm, which attempts to find the likeliest and sparsest assignment of the core predicates, once their interactions have been fixed by the proposed theory. During our algorithmic learning process, new concepts may at times have only a vague meaning, especially when they are

first proposed. Concepts that are fragmented can be unified, and concepts that are lumped together may be usefully dissociated, as learners move around theory space in ways similar to how new concepts are manipulated in both children's and scientists' theory change [26].

Returning to the overarching idea of the child as scientist, it is interesting to recall how from its inception, the study of the cognitive development of children was heavily influenced by the philosophy of science. Many researchers have found the metaphor of children as Lilliputian scientists useful and enlightening, seeing children as testing hypotheses and building structured causal models of the world, and this idea has found an exact formulation in an ideal Bayesian framework. However, neither children nor scientists are ideal, and discovering the practical learning algorithms of children may also lead us back to a better understanding of the process and dynamics of science itself as a search process.

Despite our optimism, it is important to end by stressing that our models at best only begin to capture some aspects of how children acquire their theories of the world. We agree very much with the view of [151] that the hardest aspects of the problem are as yet unaddressed by any computational account, that there are key senses in which children's learning is a kind of exploration much more intelligent and sophisticated than even a smart randomized search such as our grammar-based MCMC. How could our learning algorithms account for children's sense of curiosity, knowing when and where to look for new evidence? How do children come up with the proper interventions to unconfound concepts or properties? How can a learning algorithm know when it is on the right track, so to speak, or distinguish good bad ideas from bad bad ideas, which children seem able to do? How do pedagogy and learning from others interact with internal search dynamics - are the ideas being taught simply accepted, or do they form the seed of a new search? How can algorithmic models go

beyond the given evidence and actively explore, in the way children search for new data when appropriate? There is still much toil left – much rewarding toil, we hope – until we can say reasonably that we have found a model of children's learning, and believe it.

Chapter 5

Commonsense Reasoning About Physics and Psychology

"Common sense is the more consolidated [type of thought], because it got its innings first, and made all language into its ally." — William James, Pragmatism

---

If we could put a infant's mind directly under a lens, as one peers into a dish under a microscope, what would we see?

Perhaps one big interlocking mechanism. Or maybe a thousand tiny special-purpose cognitive organelles [170]? Probably neither, according to recent decades of research. Instead, we would see a small number of distinct regions, each with its own parcel of concepts and principles [168]. On one side of the dish, core physics. Over there a jumble of a different hue, possibly core psychology.

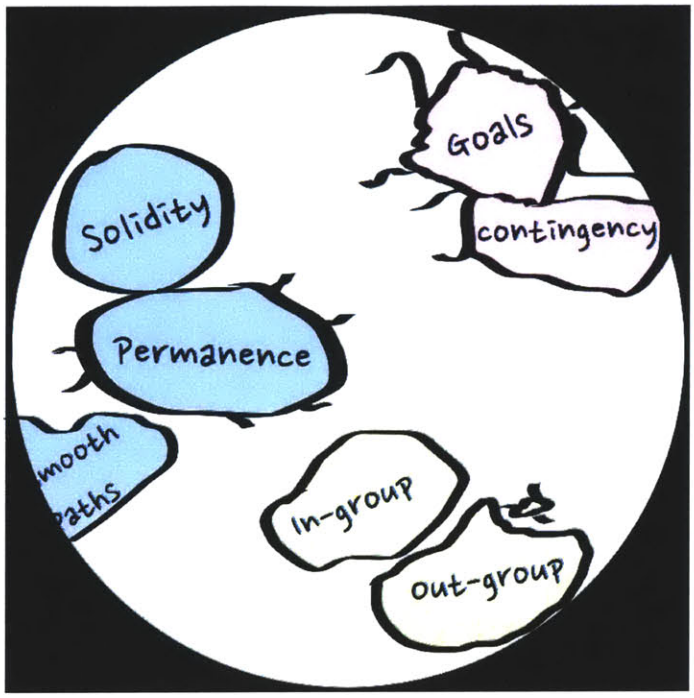


Figure 5-1: Microscope view of core domains in a mind-dish.

It seems an odd picture, beyond the oddity of imagining concepts as flattened globules drifting in space[1]. The picture is odd because it conjures up two immediate puzzles:

First, what is the inner machinery and mechanism of these globules?

Second, how do these different globules relate and communicate with one another?

To better explain the puzzle of mechanism, take for instance the glob labeled 'contingency'. It is not in doubt that infants expect agents to act contingently, and that this expectation forms part of the core understanding of agents [167]. But knowledge in such a form seems incomplete. As mentioned in the introduction, it is

[1]Researchers do *that* all the time.

more a Keplerian principle than a Newtonian explanation. How can we implement such a principle in an artificial system? Should we put in a 'contingency' detector that raises a 'surprise' flag when a violation is noticed? How would we define contingency, or agency? This question of underlying cognitive machinery applies to many of the other 'core knowledge' propositions, like 'objects should not wink out of existence' or 'objects follow smooth paths'. Chapters 2 and 3 partially addressed this puzzle by providing inner mechanisms in the form of formal models. This is not to say the models fully answer this question, but they provide the shape of a satisfactory answer.

What about the second question, how do these domains communicate with one another? One answer is that they don't. At least, not until we develop language or a language-like ability[2]. But if that's the case, how do we know which is the right core system to use when trying to make sense of a scene? And what if common-sense interpretations have to take into account several core domains? In particular, it seems that the core systems of physics and psychology have to interact frequently in order to reason in a common-sense way about scenes. Are there perhaps 'core-connections' between the core systems?

In this chapter I focus on the second puzzle. To better explain the challenges involved, I examine a "mini-world" that requires people to simultaneously use two core domains – physics and psychology – in order to make sense of it. I will set up some intuition for the problem of interest through the classic Heider and Simmel psychophysics study [79] (Section 5.1). I then propose a minimal version of Heider and Simmel's world that captures its important elements. I show this world, while shrunk and bare, can give impressions of animacy and non-animacy, tug and drag,

[2]That is one common interpretation of why young children, rats and language-hindered adults are unable to solve a spatial navigation task that requires input from non-geometric cues [188, 157]

chase and flee, push and shove, fight and run (Section 5.2). I review 'bottom-up' and 'cue-based' approaches and how they might tackle joint animacy-and-physics reasoning (Section 5.3).

I then present a generative-model, where the perception of animacy is a hypothesis-comparison process (Section 5.4). I briefly remind the reader of the formal models of intuitive physics and psychology, pointing out their similar structure and similar 'end product': entities moving over time. I propose a generative model that builds on the minimally necessary parts from both physics and psychology. The model generates observations in, and reasons about, a perceptually bare but conceptually rich domain. I end by considering some options for formalizing an 'Ur-theory' of intuitive physics and psychology (Section 5.5). The main contributions of this chapter to the thesis are: the generative model for a minimal domain that requires reasoning about both objects and agents, the minimal domain useful in itself for testing different approaches to common-sense reasoning, and the suggested formalizations for combining objects and agents.

5.1 Building Intuitions With Moving Circles

To get a better sense of the problem, consider the classic short movie used by Heider and Simmel in 1944 to examine social attributions [79]. The scenario depicts four shapes (see Fig. 5-2): a small circle, a small triangle, a large triangle and a rectangular hollow box. For a little over a minute, the circle and triangles move, rotate and change direction and speed. When asked to describe the scene, most people invoke agency, goals and social relations to explain the motion, e.g. "The girl gets worried and races from one corner to the other", "The two chase around the outside of the room together", "They finally elude him and get away", "He evidently got

"Lovers in the two-dimensional world, no doubt; little triangle number-two and sweet circle"
– A participant

banged around and is still weak". When asked to the describe the personality of the objects people are also consistent. The large triangle is seen as "a bully", "a villain", "irritable", "angry", "pugnacious" and so on.[3]

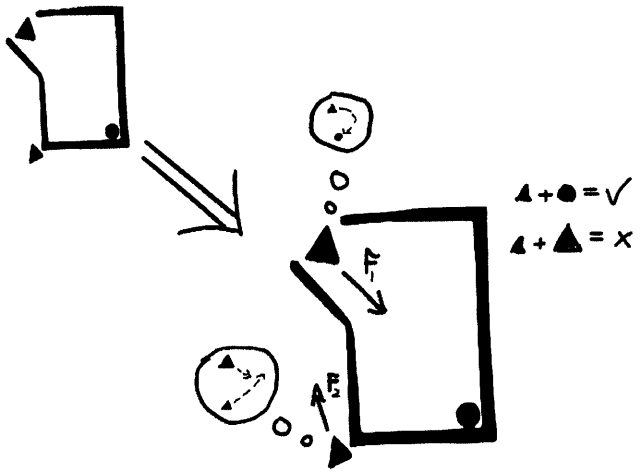


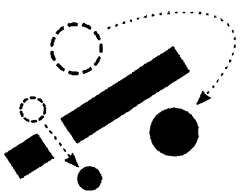
Figure 5-2: Frame from the classic Heider and Simmel stimuli in top left corner. Lower right corner is caricature of some of the 'unobserved' information that people read off the stimuli: agency, social relations and physics. Original movie is embedded, click the picture to view.

Importantly, hardly anyone describes the scene as 'The triangle is moving downwards at about 2 centimeters per second, it changes direction and is heading towards the inner wall with reduced speed...' although this description is as valid, and in some ways more accurate[4].

[3]Not *every* video of moving shapes elicits consistent judgments. Heider and Simmel also ran the movie in reverse and recorded much greater variation, as the following response shows: "Man (T) finds himself in chaos, which finally resolves itself into a sort of cell representing Fate. He is able to free himself (but only temporarily), when Woman (c) accompanied by Evil (t) comes upon him, and disrupts his momentary peace. He feels called upon to rescue her, but Evil imprisons them both by Fate, from which Man escapes, leaving the woman there for safe-keeping. He at first seems to vanquish Evil, but Woman comes into the picture again and again disrupts Man. She goes off with Evil, as he seems the winner of the struggle, and Man, not understanding her, himself, or anything, resigns himself to Fate."

[4]In the original study only 3 out of 78 participants did not use animacy to explain movement.

This well-known example :



Not just animacy, physics too While people use intuitive psychology to explain the movements of the shapes, they still need to posit something about the physics of the situation for the common-sense explanation to make sense. People apparently assume that agents are solids and not inter-penetrable, that movement incurs a cost and requires some exertion, that movement happens locally, that collision involves a transfer of momentum, that the 'door' will not move on its own and requires effort to open, that the walls constrained the movement of the agents and so could 'trap' the small agent, and so on. So while the stimuli is often touted as a case of animacy and social attribution, it is more correctly described as an interplay between animacy and physics[5].

A lot from a little The stimuli has only three simple geometric shapes, as well as one 'wall-shape'. The state of the shapes is characterized by their x-y coordinates and their orientation. The movie is approximately 70 seconds long, with 'state-changes' (rotations or movements) happening at most twice per second. So, a complete physical characterization of the state of this 'world' evolving over time can be represented as a matrix of $9x140$ (position and orientation for 3 shapes over 140 steps), and some additional data about the shape dimensions and wall positions. One could imagine similar low-dimensional stimuli involving more or less shapes and somewhat longer or shorter scenarios, but all within the same general category of 'Heider-and-Simmel-Stimuli' (HSS). So, HSS seem like a tractable target for the computational modeling of how people

This data is pooled across conditions in which participants were asked to describe the scene without overtly being told to use agent-like explanations [79].

[5] [33] explored how constraints affect the interpretation of animate behavior, showing how physical objects (like walls) are necessary for perceptions of goals (like chasing) in infants.

attribute intent and relation, goal and effort, mass and emotion.

5.2 Lineland, a minimal Heider-and-Simmel world

5.2.1 Why not model the original Heider-and-Simmel world directly?

Is an input of $9x140$ tractable, though? To answer that we must consider the size of the input space itself. In the same way that $30x30$ pixel squares can be used to tractably study object recognition, but also produce a frightening number of possible inputs[6]. A large input space means that planning algorithms, necessary for models of intuitive psychology, might falter. I'll sketch the planning space, and then introduce a more minimal HSS which is even more tractable than the original HSS.

Assuming that something like physical dynamics governs an HSS, not every trajectory is possible. That is, an object cannot suddenly radically change its (x, y) position nor enact a huge force of any size in any direction. Assume that at any point in time an agent can take an action out of a set A, which could include rotating clockwise and counter-clockwise, or moving in some radial direction with some force. For the sake of simplicity, we assume that there are 4 rotating actions: Big rotate clockwise, small rotate clockwise, big rotate counter-clockwise, small rotate counter-clockwise. Further assume that there are 16 moving actions, crossing compass directions with movement step-size (big/small step northeast, big/small step north, big step northwest, etc.). If we include 'do nothing', we have 21 possible actions per entity, at every time-step. Thus we have a space of 63^{140} joint-plans to consider. That's a lot.

[6] 2^{900} assuming pixels are black or white.

RRT searching a 2D space

One could try to get around the size of the input space by considering sampling-based planning algorithms, such as Monte-Carlo-tree-search [29, 19] or Rapid Random Trees (RRT) [105]. RRTs are particularly suited for this domain, being useful for cases where there is a finite set of actions but a continuous state-space governed by force dynamics. Such dynamic planning algorithms are at the forefront of current planning research, and are difficult to invert for inverse planning. The point of this chapter is not to wrestle with the implementation of a suitable dynamic planner, but to consider a minimal version of the problem and see what can be said about the necessary components of even a simple generative reasoning model.

A minimal HSS should be tractable, but also retain the interesting properties of the original HSS. This brings us to Lineland.

5.2.2 Introducing Lineland

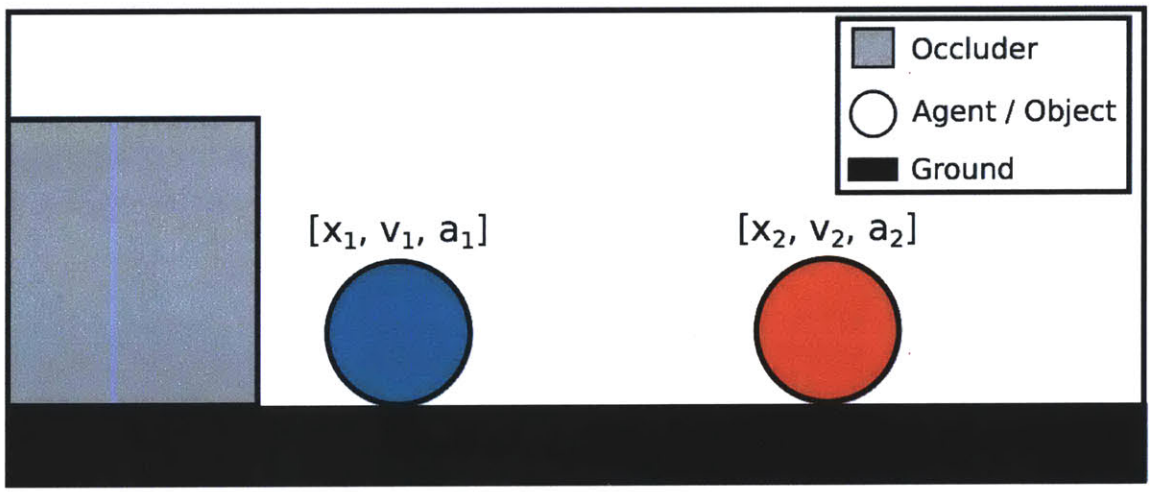


Figure 5-3: Depiction of Lineland, including visible elements and properties. Circular entities all share the same y-position and can only move along the x-axis, thus their state is fully specified by a single number x.

In Lineland there are entities (circles), occluders (squares) and the ground (ground). Entities can only move along a one-dimensional line, and they all share the same y-position. Lineland is also discrete in space and time, and entities cover an integer distance at each time step. Since the circles have no orientation, the state matrix defining a 'scenario' in Lineland is $2 \times N$, where N is the number of discrete time steps allowed.

"I had retired to rest with an unsolved problem in my mind. In the night I had a dream."
– The Square imagines Lineland

The original HSS included a 'room' object and 3 entities moving and spinning in a 2D environment. In this restricted Lineland we have 2 objects moving left and right along a single dimension[7]. The set of possible 'trajectories' to consider in Lineland will depend on the set of possible actions each entity can take, but it is clearly a small subset of the original HSS space.

5.2.3 Scenarios in Lineland

Can such a world still support social and physical inferences? Yes, as I'll show below. I discuss 12 different stimuli in a restricted version of Lineland, where scenarios unfolded over just 8 time-steps. These stimuli were shown to 100 participants on Mechanical Turk (59 men and 41 women). Participants were told that they would see short snippets of movies from 'Lineland', movies that depict colored shapes moving around[8]. They were informed that some shapes are 'people', while others are 'objects'. Participants were asked to:

1. Describe in free-form what happened in the movie.

[7]In the 1884 book Flatland [1], the Square protagonist actually considers a one dimensional world called Lineland, but there objects are lines and dots without area, quite ill-suited for our purpose. Still, the name seems apt.

[8]The full experiment can be seen here: `http://www.surveygizmo.com/s3/1896329/Lineland-2`

2. Pick any of the following labels that apply to the movie: Chasing, fighting, bumping, enemies, friends, attracting, repelling, dragging, resisting, pushing, sliding, fleeing.

3. Decide for each shape whether it is a 'person' or an 'object', and if it is a person to choose which of the following goals best describes it: Move itself to the right, move itself to the left, move the other shape to the right, move the other shape to the left, help the other shape, hinder the other shape, get close to the other shape, run away from the other shape. ...

I present the scenarios verbally and with accompanying figures. Each figure includes a schematic of the scenario, participants' animacy ratings and a histogram of the labels participants endorsed. The free form responses can be found at `http://www.mit.edu/~tomeru/lineland/lineLandResponses.txt`. Note that of 100 participants, 30 judged none of the entities as a 'Person' in any scenario (Fig. 5-4), and so any variability in the animacy ratings is going to be due to the other participants. Given that, the animacy ratings omit these 30 constant participants[9].

"It looks like a billiards game" – Participant

Launching, equal mass: Fig. 5-5(a) shows a classic Michotte-like launching [119]. The Blue Circle (BC) begins at uniform speed, stops upon contact with Red Circle (RC), which begins motion at the same uniform speed. Variants of this trajectory have been used to explore impressions of causality, varying the temporal and spatial gap between objects [148]. It's not my intention to tread over this well-trod area again, just to note that this impression is possible.

"Blue slides into Red but Red is a heavy load and just repels Blue." – Participant

Launching with different mass: Fig. 5-5(b) shows BC moving at uniform speed towards RC as before, but this time upon contact it moves back to the left

[9] It's possible those participants were consistently being lazy, as choosing an entity as a 'Person' means choosing their goal

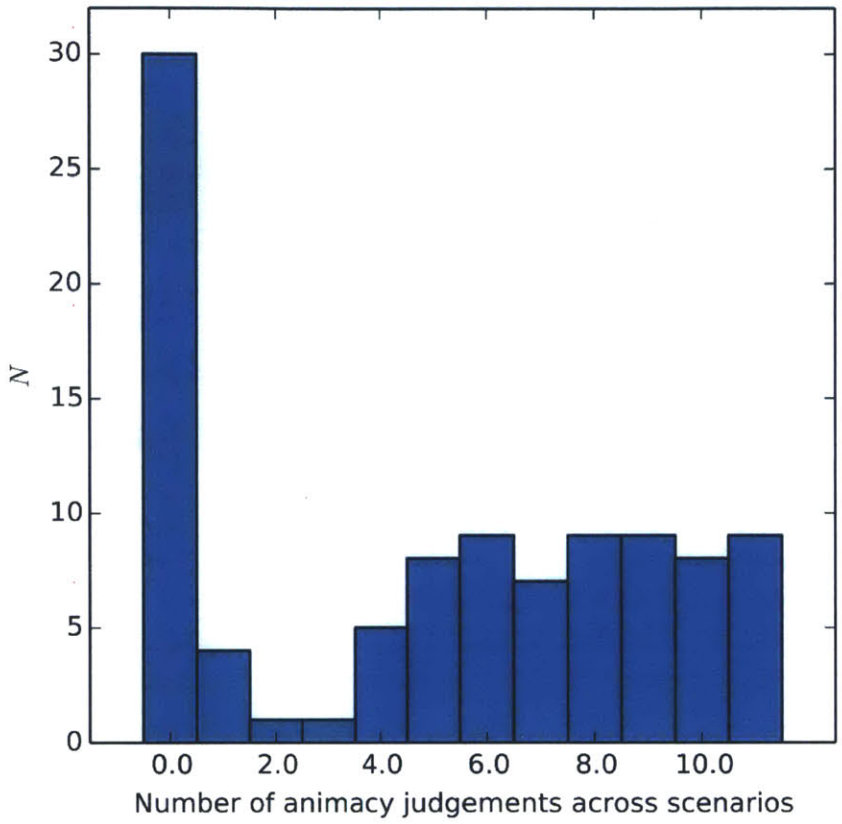
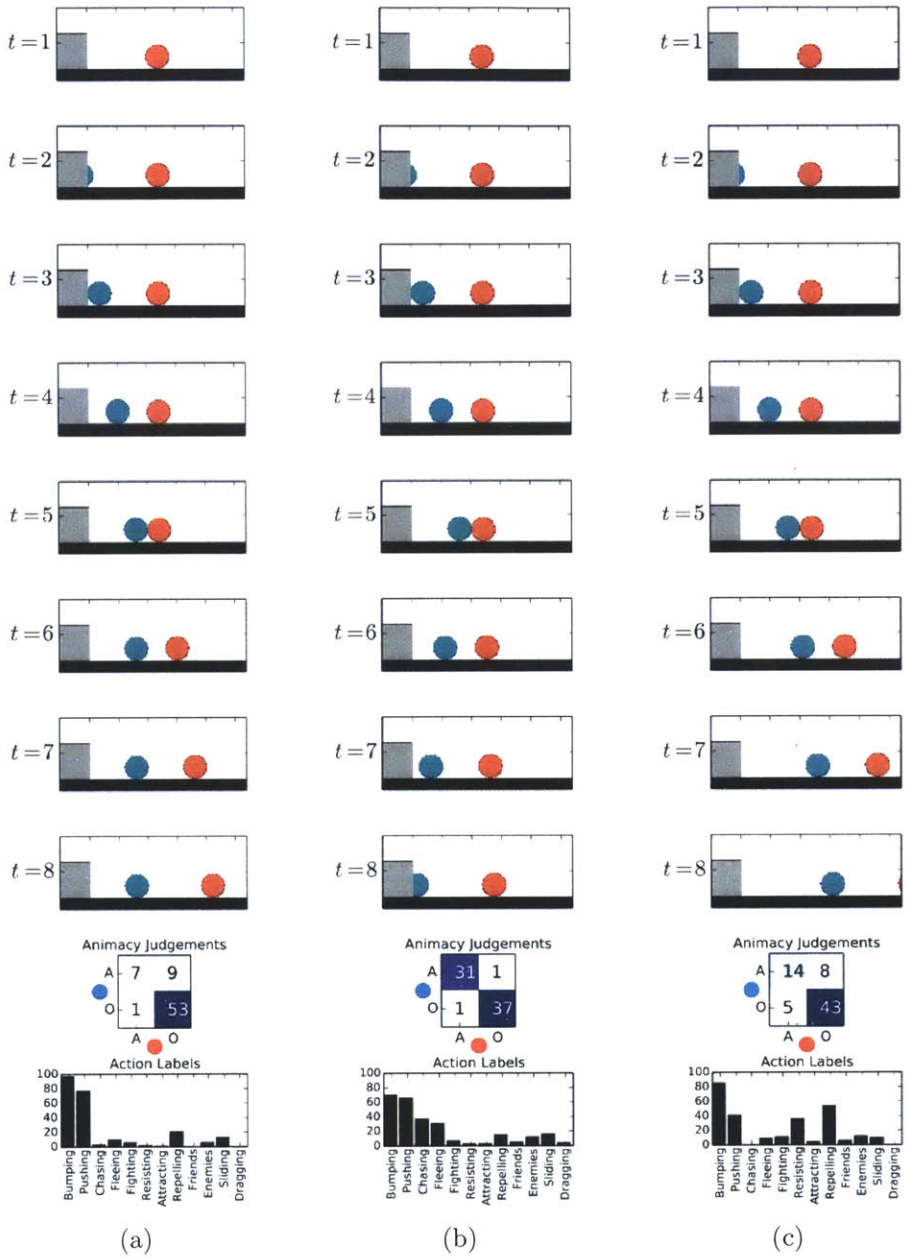


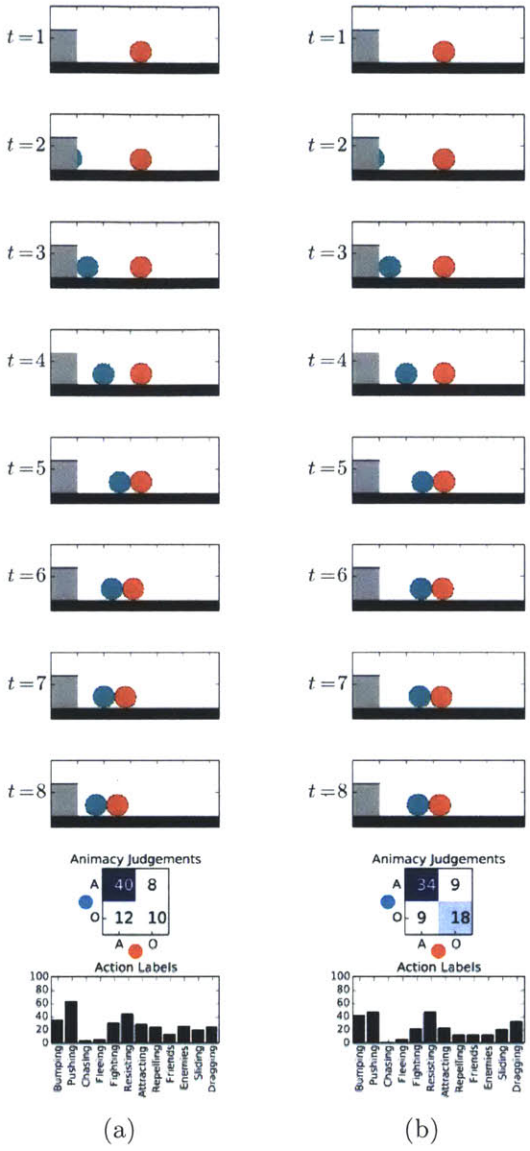
Figure 5-4: Histogram of animacy ratings across scenarios. For example, in 10 out of the 12 scenarios, 7 people judged at least one of the entities as animate. 30 people judged none of the entities as 'person' in any scenario.

at reduced uniform speed, while RC begins moving to the right with low uniform velocity. Fig. 5-5(b) shows BC moving with uniform speed into RC, and continuing at reduced speed to the right, while RC begins moving to the right at a greater speed than the BC had initially. The impression is meant to be that of launching, similar to but with a sense that the masses of RC and BC are different. In Fig. 5-5(b) the impression is that $BC < RC$, while in 5-5(c) the impression is that $BC > RC$.



(a) (b) (c)

Figure 5-5: Static images of dynamic sequence giving impression of BC launching RC, with (a) $mass_{BC} = mass_{RC}$, (b) $mass_{BC} < mass_{RC}$, (c) $mass_{BC} > mass_{RC}$



(a) (b)

Figure 5-6: Static images of dynamic sequence potentially giving impression of BC dragging RC, or RC pushing back BC. Depending on the interpretation (dragging or pushback), either $mass_{RC}$ or $mass_{BC}$ are smaller in (a) than in (b).

Dragging/Pushback: Fig. 5-6(a) and (b) show snapshots of a stimuli intended to more directly evoke a sense of animacy. BC approaches RC, comes into contact with it and then reverses direction, followed this time by RC. It seems that either BC is coming up to RC and "dragging" it back to the left, or RC is pushing BC back after the two encounter. Participants reported seeing either one or the other, or both. The two scenarios should also evoke different senses of mass compared to one another, but the exact inference should depend on the interpretation. If BC is being pushed back by RC, than BC should appear heavier in Fig. 5-6(b) compared to (a). If BC is seen as dragging RC then it is RC that should appear heavier.

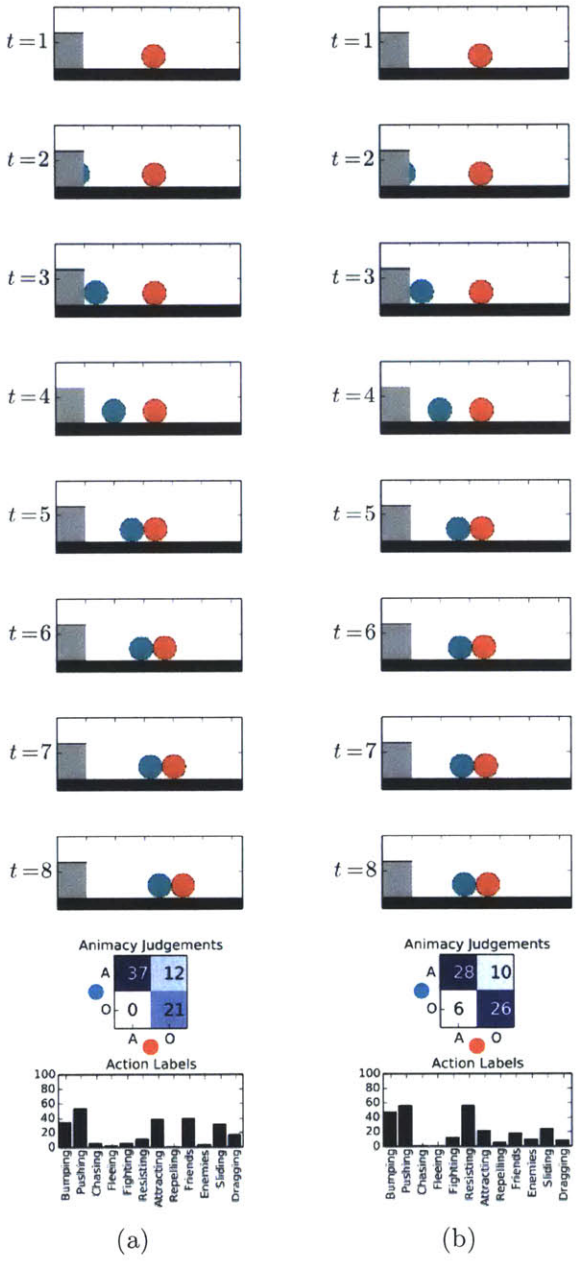
"blue meets red and takes him home" –Participant

Pushing: The following stimuli, shown in snapshots in Fig. 5-7(a) and (b), is intended to be ambiguous between animacy and non-animacy. BC moves towards RC, comes into contact with it, and then continues to the right, staying in contact with RC who is also now moving to the right. The impression is that BC is possibly pushing RC, whose mass in (b) is greater than its mass in (a). Alternatively this could be seen as a plastic collision, with both circles being inanimate objects and BC having started with some initial velocity.

"The blue ball rolled into the red ball and pushed it." –Participant

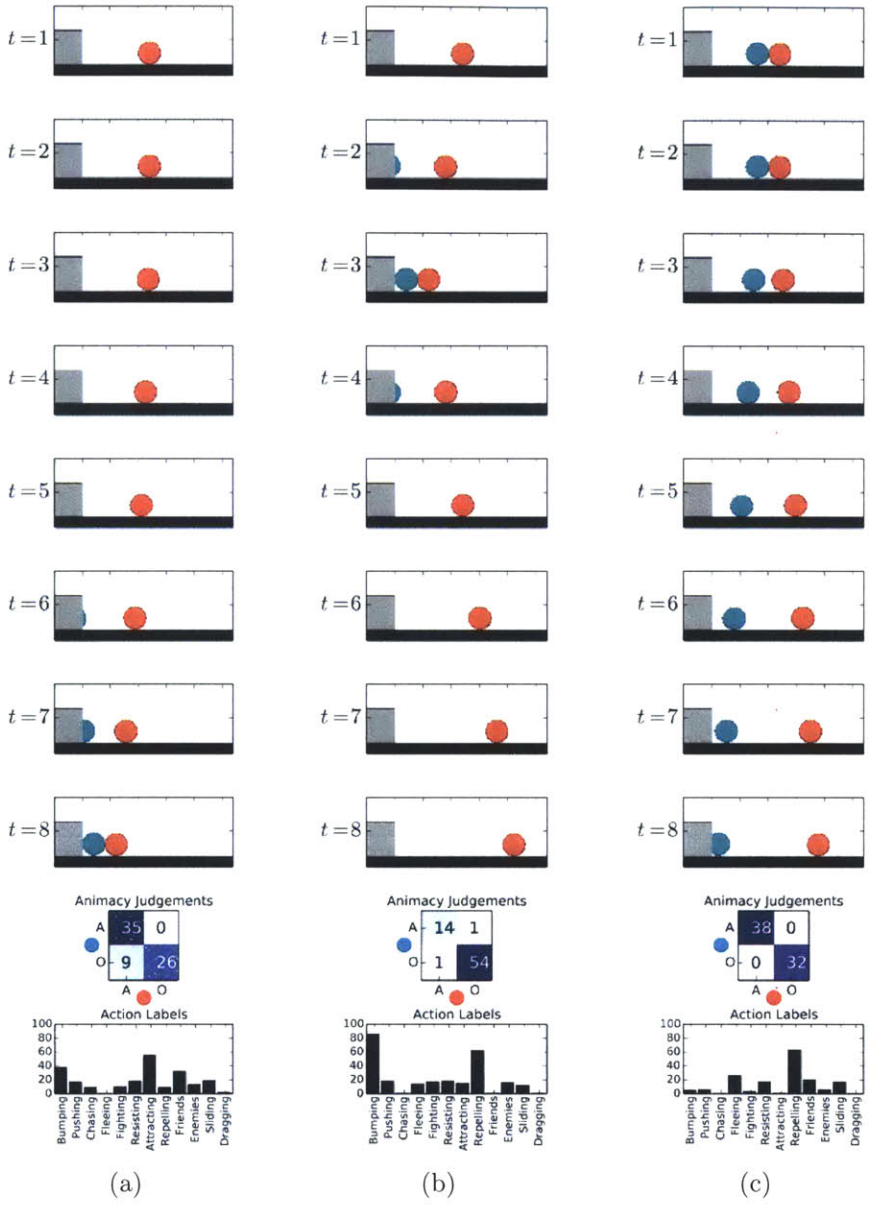
Attracting and repelling: Another set of ambiguous stimuli involve RC and BC both moving towards or away from one another. Such stimuli, showing in Fig. 5-8(a-c) can give impressions of attraction (a) and repulsion (c), although they can also be interpreted as the objects starting with some initial velocity towards one another and bouncing off. Attraction and repulsion could be interpreted as a physical force pushing and pulling at the objects, or as the objects having some goal of being closer or farther from one another. The impression of acceleration and deceleration may also play a part here. Consider how uniform velocity and then stopping may be seen as more 'intentional' attraction than an acceleration, which is more in line with an attractive force dependent on distance. The modal labels for (a) were 'attraction'

"I feel like they are coming together as friends or like a magnet and being attracted to one another." –Participant



(a) (b)

Figure 5-7: Static images of dynamic sequence potentially giving impression of RC pushing BC with $mass_{RC}$ in (a) being smaller than that in (b).



(a) (b) (c)

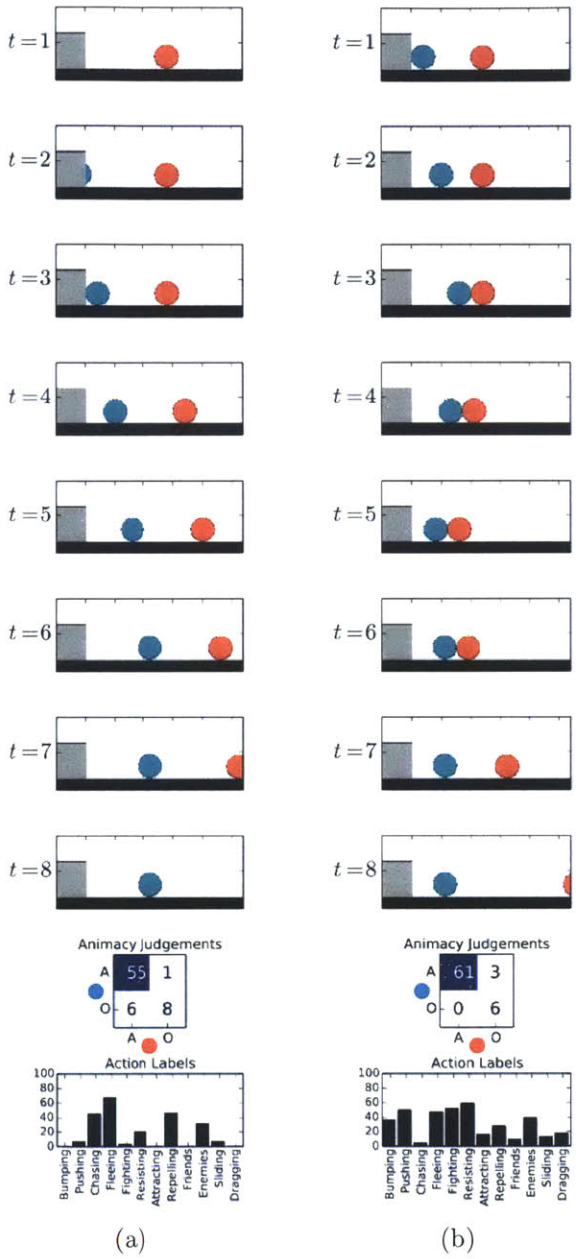
Figure 5-8: Static images of dynamic sequence with *RC* and *BC* both moving towards or away from one another, potentially giving impressions (a) attractive forces, (b) physical bouncing and (c) repelling forces.

and 'friends', for (b) it was 'bumping', and for (c) 'repelling' and 'fleeing'.

Chasing, fleeing and resisting: Fig.5-9(a) and (b) show stimuli designed to evoke a sense of animacy for both entities, not just *BC*. In Fig.5-9(a) *BC* beings to move towards *RC*, which itself begins moving away from *BC* before contact is achieved. *RC* accelerates and disappears from view. This stimuli evoked strong senses of 'chasing' and 'fleeing' participants, although some report seeing 'repulsion' or 'pushing without touch'. In Fig.5-9(b) *BC* makes contact with *RC* and then begins moving back to its original location, with *RC* moving in that direction too. However, the two both shift direction, and then a gap is formed between them as *RC* accelerates to the right. Some participants reported this as a case of several shoves, while others described 'a fight', 'a power struggle' or 'a mugging'.

"Fight!" -Participant

Many other scenarios are possible. I focused on this subset to show how even in this simple world people can have different possible impressions of physical properties, goals, agent-object interactions and agent-agent interactions. A "whole-scene" interpretation relies on reasoning about the physical, psychological and social domain at the same time, and as in the original HSS, if we had a model that took these displays as input and produced common-sense explanations as output, we will have gone a long way. But what is the right way to construct such a model? One temptation is to focus on the 'simple' aspect of these simple trajectories, and conclude that a kind of motion-feature library could be built of these space-time curves, used for classification without the need to bring in complicated mental models. This has certainly been the direction in some branches of development and perceptual research, which I consider in the next section.



(a) (b)

Figure 5-9: Static images of dynamic sequence potentially giving impression of (a) *BC* chasing after a fleeing *RC* and (b) *BC* struggling with *RC* which then flees.

5.3 Perceptual-Cue Classification of Objects and Agents (and Why It Probably Won't Work as a Standalone)

5.3.1 The cue-based tradition

As discussed in Chapters 1-3, there is a long history in psychology of trying to find perceptual cues for the attribution of physical and psychological states. Going back at least to Michotte's observation that people see in situations much more than is directly perceptually present, from 'causality' to things such as 'drinking', the view was that a scene had certain properties that were picked up by perceptual detectors that triggered a psychological state. For example, a detector for for 'causal launching' might take the form "*IF* an object A approaches a stationary object B, *AND* upon contact object B moves, *THEN* object A is perceived to 'cause' B to move" [119]. Such an observation is then embellished on by countless studies of the exact conditions which ignite the 'launching trigger', varying velocities, spatial and temporal gaps, and so on. The perception of more specific physical properties (such as mass) were also suggested to come about through cue-heuristics (such as relative velocity) [140, 55].

The perception-based view was also extended to social and psychological attributions. Again going back at least to Michotte [119] but continuing up to the present [148, 181, 47], the idea is that the trajectory of an object has certain properties that are picked up by fast perceptual detectors for distinguishing simple goals ("it wants to get there"), social goals ('dancing', 'chasing', 'hunting', etc.), relations ('friends', 'enemies', 'lovers', etc.) and basic distinctions such as animacy/inanimacy. As one

proposal put it, if an object suddenly changes its velocity without any obvious external obstacle or constraint, that is a strong cue for animacy [181]. Additional detectors refine the perception. For example, one detector might detect animacy, and then if two animate objects take part in a particularly tight spatio-temporal trajectory, they might be seen as 'chasing' or 'dancing' [148]. The hope of these approaches is that one can build up a 'grammar' or 'library' or 'set of components' of simple motion-paths [147], which either interact to convey the meaning of more complicated paths, or simply compete for dominance of the explanation. The general research program is to analyze simple scenes and hunt for the minimal 'cues' or 'features' that distinguish mental attribution. As one example, Blythe has suggested that seven motion cues are sufficient for distinguishing animacy from non-animacy as well as the intention of animate agents in several tasks [16]. Researchers in the cue-based tradition usually suggest these cues are either built-in or early-emerging, based on work with human infants [148, 33], primates [183] and across cultural groups [11].

Researchers in development also hunt after 'cues' for physical causality and animacy (also referred to as 'social causality') [146, 145, 143]. There is an ongoing debate about how much experience matters in detecting animacy, but work with newborns [143] and control-reared chicks [113, 135] found that at least some 'cue-detection' is present at birth in vertebrates. These innate cues include 'physical causality' detectors for Michotte-like causality, 'animate' detectors for self-propelled motion, and 'life' detectors for (upright oriented) biological-like motion [143]. This view leads to fine-grained experiments regarding the exact relevant cues, and to debates about whether it is self-propelled motion that is the primal cue, or perhaps contingent interaction [145], etc.

What would the feature-based approach make of Lineland? It might begin with an analysis of the curve of the 'world-line' described by each entity, and the distance

between them, or it might show these stimuli to infants, children and adults to try and tease out the relevant cues for judgments like 'power struggle'. It might even succeed in identifying some useful cues for the stimuli presented.

Here, then, is one broad characterization of what takes place during infancy and development according to this view (see also Fig.5-10): A baby views a certain scene. Static and kinetic properties of the scene trigger innate or early-developing detectors that "push" the processing of the scene into one of the domains of core knowledge. For example, self-propelled motion might trigger 'animate' detectors, which relay the processing to the core agency system. Once the scene is classified as belonging to certain a core knowledge domains, it is subject to innate or early-developing expectations. For example, an animate agent is expected to propel itself, resist forces and exhibit preferences, etc. [168]. Further detectors are called on to refine the classification, either in a cascading hierarchy (self-motion analyzer detects animacy $\rightarrow$ relative velocity analyzer detects social interaction $\rightarrow$ vorticity analyzer detects courtship $\rightarrow \ldots$), or just all 'fire' at once. Development and learning is then the training of further classifiers / decision-trees to guide expectations about particular scenes [6]. For example, children might learn that in a scene classified as 'inanimate object stability scene', it is important to pay attention to the variable 'distance from edge', and that there is a rule such that IF an object is inanimate and is a certain distance from the edge of the table, $THEN$ it should fall [6].

This view is incomplete. I do not mean that some cues and heuristics are not real or useful[10], but it seems hopeless to think that a long list of cues and features can make something that has as output the reports of people upon seeing the Heider

[10]As discussed in Chapters 2 and 3, these cues might act as lens that quickly focus attention on small parts of a hypothesis space, or they might act as 'hooks' that a pre-existing conceptual framework can use in order to latch onto parts of a perceptual scene [185].

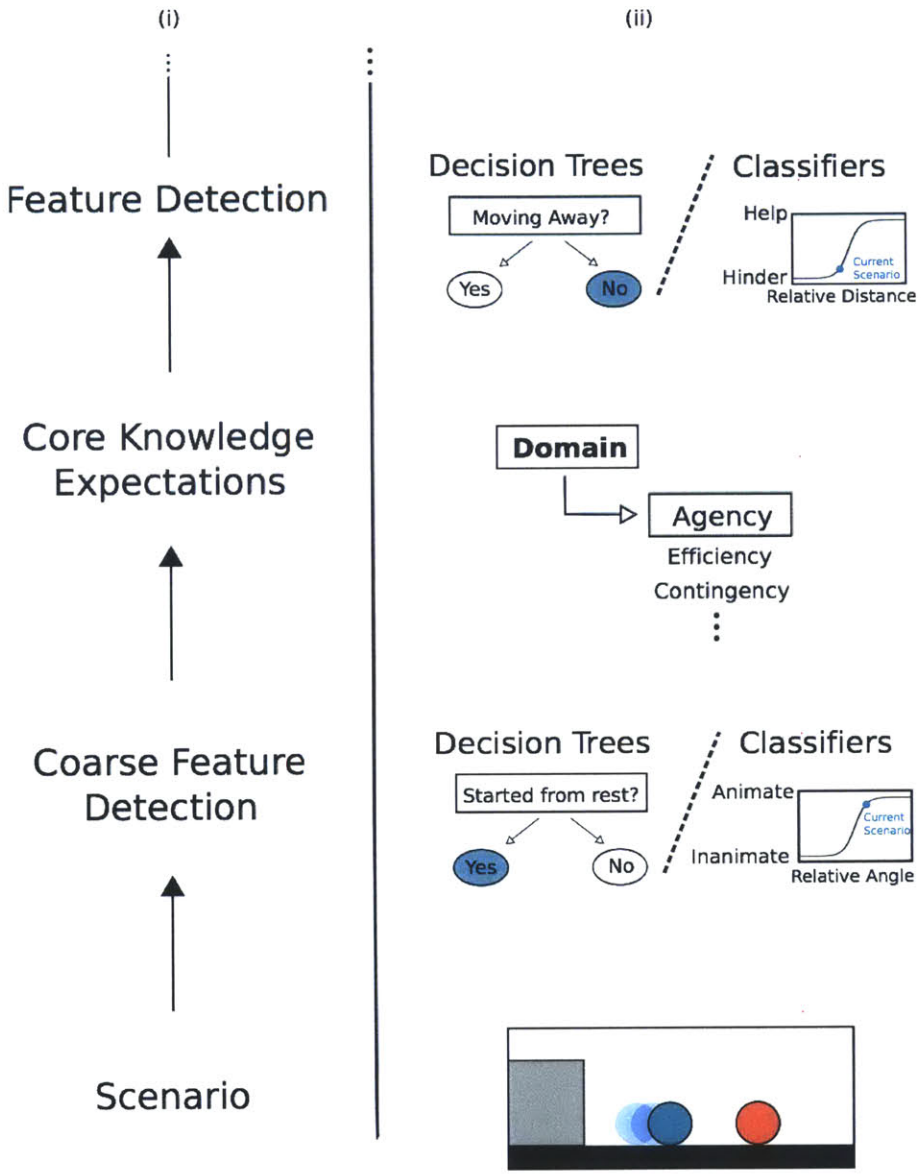


Figure 5-10: Characterization of feature-based approach to core knowledge: (i) The general progression is from a perceptual scenario going through finer and finer classification using different features. (ii) Applying the stages on the left-column to a specific example.

and Simmel stimuli. And stimuli that are *too* simple – like using a single object on a 2D plane and varying its acceleration [181] – might not be a way of finding minimal cues of animacy, but rather a pathological case where theory-based attribution is stretched to its limits.

Scholl and colleagues, in particular, argue against the involvement of 'higher-level' cognition in making fast, consistent and automatic social and physical attributions. As they put it, 'The perception of animacy is more akin to the perception of depth and color in a painting than to the perception of sarcasm or irony in a painting" [147]. But in between these two extremes there is still a wide range, one that is occupied among other things by our core knowledge - rapid and difficult to penetrate, but still conceptual to a certain degree, and possibly captured by theory-like generative models such as the ones described in this thesis.

In the next section, I propose a different approach. When distinguishing between 'animate' and 'inanimate', we must first describe what our theories of the physical and psychological world are, and how they relate. The perception of animacy can then be seen as a form of *hypothesis-testing* between these competing theories. Certain aspects of the motion paths might emerge as particularly diagnostic for such a calculation, and the most general ones might turn out to be useful cues to build into any intelligent system. But those cues are in the service of the more theoretical-based distinction, not the distinction in and of itself.

5.4 A Joint Model for Reasoning About Physics and Psychology

5.4.1 General Considerations

In the previous chapters I considered models of physical and psychological reasoning in isolation, but the two seem linked on a fundamental level. I consider the parallels between the two domains and their formalization, contrasted with other domains.

The elementary building blocks of the **intuitive physics** model are *entities* with *properties* that update according to *forces*. Some of the properties are assumed "privileged" - they are shared by all entities. *Entities* are further divided into *dynamic entities* and *static entities*, dynamic entities being the ones expected to move and so having the privileged property mass. The *forces* act on dynamic entities and update their accelerations under a general assumption of noisy Newtonian dynamics, where $a = \frac{F}{m} + \epsilon$.

The elementary building blocks of the **intuitive psychology** model are *agents* with *utilities*, with different possible *states* updated according to *actions*. Agents differ in their recursive reasoning abilities and relations, and utilities can be divided according to whether they are 'about' states of the world, or other utilities. Agents generate according to their utility functions, under a general assumption of rational planning, and these actions update the state of the agents.

Using these building blocks, each of the frameworks construct their hierarchy, each level bringing them closer to specific observable scenarios. The intuitive physics model specifies particular forces (springs, attraction, ...) and entities (pucks, blocks, magnets, ...). The intuitive psychology model specifies particular utilities (location goals, social goals, ...) and agents (evil, good, recursively reasoning, ...). The intuitive

physics model grounds out as physical objects moving along observable paths. The intuitive psychology model grounds out as agents taking a series of observable actions in pursuit of goals.

Each of these frameworks represents a unification of principles within separate core domains [168]. Even though their elementary building blocks appear qualitatively different, these formalizations show that the general structure of the two domains is similar: Both frameworks proceed along similar lines by going from abstract principles to grounding out in perception. Both frameworks support similar inference mechanisms and learning at all levels of the hierarchal model, and they can both be applied to a potentially infinite number of real or fictional scenarios. Both frameworks are concerned with entities producing some observable trajectories. Finally, both frameworks produce scenarios that can end up looking similar: *entities moving over time.*

So, there are similarities in both the functional form and the end-products of these core domains. But could such an argument be made for any pair of core domains? Not necessarily. Consider for example the core domain of number, and a recent model of number word acquisition suggested by Piantadosi et al. [130], which is also based on probabilistic generative programs. This model assumes a small number of primitives for testing the cardinality of sets (e.g. *singleton?, doubleton?*), the ability to perform simple set manipulations (e.g. *set-difference*), simple logical operators, recursion, and the combination of primitives. Using these primitives and given the right data, the Piantadosi model can go through the stages of number acquisition, moving from a 1-knower (someone who can pick out a set of size 1, but nothing else) to a Cardinal-Principle-knower (someone who can pick out sets of any size) ([26]. Some of these required primitives (like recursion and logical operations) might exist in other core domains, though it is an open question whether the primitives are part

of a shared ability (a general 'and' operation) or exist independently in each domain (and-number, and-agent, and-object and so on). Still, the majority of the relevant primitives are not shared at an abstract level with psychology and physics. The hierarchical structure of the generative program does not match the other domains, nor does it lead to data that explains the motion of entities. The whole domain might exist as a sort of 'function-call' within a task in the other domains when needing to keep track of the entities, but it is not a competing hypothesis for the explanation of the data per se.

5.4.2 The Formalization of Lineland

Based on the parallels described in the previous section, the minimal things that need to happen in order to tie the models of physics and objects are these: First, the transition function $T(S, A) -> P(S')$ needs to be informed by a physics-engine, rather than being built in by hand as no different from a general constraint[11]. Second, the inference must consider both physical forces and agents as potential explanations of motion. Third, the actions of agents must be 'physically appropriate'. That is, the actions need to be made in a way that the transition-function-physics-engine can accept as input. Applying these ideas to Lineland leads to the following model, depicted in Fig. 5-11

Briefly summarizing, the generative model of Lineland works as follows: A number of entities are chosen; all entities are assigned agenthood/objecthood and physical properties; the entities assigned agenthood are also assigned a goal; the world is assigned general force dynamics; the entities are given initial conditions (posi-

[11]The reason an agent in the grid-worlds of Chapter 2 could not go into the black squares is not because they are considered 'wall objects' with the appropriate collision dynamics, but because the transition matrix was hard-coded so that agents cannot go on black squares.

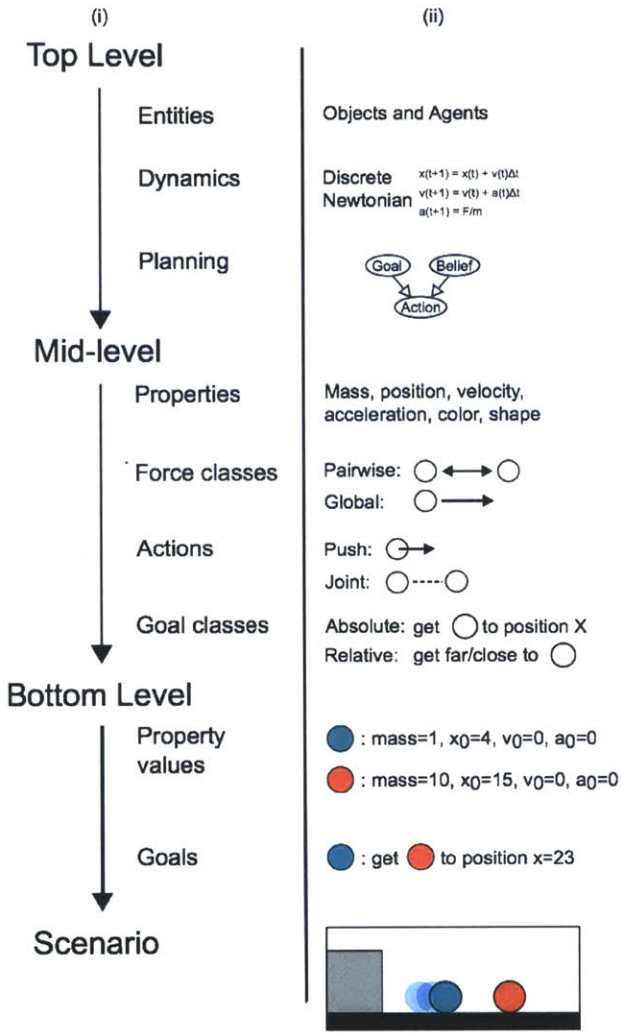


Figure 5-11: A generative approach to joint physical and psychological reasoning in Lineland: (i) The general progression is from top level assumptions about dynamics and agency in general, through finer and finer specification of what agency and physics is like, bottoming in an observable scenario (ii) Applying the stages on the left-column to Lineland.

tion, velocity, acceleration); the agents select a plan of forces; the scenario plays out according to the discrete physics engine combined with the actions of the agents;

perceptual noise is added to the scenario.

Following is a description of the model is more depth:

Top-level, entities: The possible entities of interest in Lineland are objects and agents. Agents and objects both have similar physical properties, some of which are observable (e.g. shape, position, velocity) and some of which are unobservable (e.g. mass). The properties of both agents and objects are updated by the transition function T, which is now equivalent to a physics-engine "step". A "step" operation takes the state of the world as input, including constraints and forces, and updates the accelerations, velocities and positions of all objects using a Newtonian-like dynamics [28].

Top level, discrete dynamics: The 2-dimensional world is divided by a grid into cells (similar to Chapter 2, except entities can now take up several grid-cells). An agent or object can only move in cells, which I will measure in units of "cell" or c. Time is also discrete. So, the dynamics are quite simple:

$$x(t) = x(t-1) + v(t-1) * \Delta t \tag{5.1}$$

$$v(t) = v(t-1) + a(t-1) * \Delta t \tag{5.2}$$

$$a(t) = \frac{F}{m} \tag{5.3}$$

Where the total force acting on a particular entity F_e is the vector-sum of all the different forces acting on it:

$$\forall e \vec{F}_e = \sum_{i=1}^{N} \vec{f}_e^{\,i} \tag{5.4}$$

Since Lineland is one-dimensional, I will omit the top-arrow vector notation and adopt the convention that positive numbers indicate forces, accelerations and velocities 'pointing' to the right, and negative ones 'point' to the left.

I set $\Delta t = 1$ in arbitrary units called "Steps" or s. So velocity will be in units of $\frac{c}{s}$. Acceleration will be in units $\frac{c}{s^2}$. Collisions are handled by considering the pre-collision velocity of entities and their masses, and solving for a one-dimensional collision[12].

Top level, actions as forces: Just as in the 'inverse planning' framework [8], at any time step every agent takes an action $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of all possible actions. In order for these actions to be compatible with the physics-engine transition function, actions are local forces that an agent can generate on itself and its immediate surrounding. Such an agent, even without any goal, can be seen as capturing the notion of *mechanical agency*, in that it is capable of self-propelled motion and resisting forces acting upon it [26, 160]. Beyond this, agents are also able to create and annul constraints between themselves and nearby entities (commonly referred to as 'joints' in the physics planning world [28]).

Top level, planning: Again as in the 'inverse planning' framework, agents take their actions $a \in \mathcal{A}$ at any given time step in order to maximize their utility or satisfy a goal, constrained by their beliefs about the world. Goals and utilities can be defined over the state of the world itself, or in relation to the utilities and goals of other agents [186]. The exact nature of the planning mechanism is less important here, whether it is a Markov Decision Process (MDP) [9, 8], an RRT [105], or something else. It suffices that there be *some* forward-mechanism that goes from probability

[12]Nearly all popular physics engines include a 'collision-detection' module that carries out pre- and post-collision computations, although the exact method for resolving collisions differs between one engine and the next. See also the Afterthought to Chapter 3 regarding the cognitive implications of this ubiquitous engineering notion.

distributions over states and utilities to actions. We can approximate such a planner through a procedure described later in the next section.

Mid-level, space of possible forces and physical properties: Similar to Chapter 3, I will restrict the forces to be either 'global' or 'pairwise'. Global forces act on all entities either in the positive or negative direction (to the right or to the left, when considering a one-dimensional line). Pairwise forces are either attractive or repulsive. Directly observable properties include color, shape and position. Mass is an unobservable physical property which I restrict to be 1 or 10. One can imagine making friction a possible unobservable property similar to Chapter 3, but for the purposes of this chapter it is possible to ignore it.

Mid-level, space of possible actions: I will restrict the set of actions to be $\mathcal{A} = \{\pm K; \pm joint\}$. $K = 0, 2, 3$ are the possible magnitudes of the force, where 0 indicates a non-action. A $+joint$ means the agent creates a joint-constraint with a nearby entity, while $-joint$ is a removal of such a joint. This means that given N time-steps in a scenario, there are 7^N possible trajectories for any setting of the initial conditions and physical properties.

Mid-level, space of possible goals: I will restrict the set of goals to be of the form 'Get entity E as close as possible to position P', where a position could be absolute or in relation to another entity. So, a goal for $Agent_1$ could be to get itself to a particular world position, or get $Agent_2/Object_2$ to a particular world position, or get as close/far away as possible to/from $Agent_2/Object_2$.

Bottom level: Entities are assigned initial properties (positions, velocities, accelerations, mass). Agent entities are assigned particular goals, and specific force values are sampled.

Observable data: Based on these initial conditions and in combination with the force-actions of the agents, the discrete physics engine described above simulates

forward a trajectory for some number of steps. Finally, Gaussian perceptual noise is placed on each point along the trajectory.

5.4.3 Planning in Lineland, utilities and costs

As mentioned in the previous sections, there are several ways to solve the planning problem in Lineland, including Multi-agent MDPs and RRTs, but the details of the specific algorithm matter less. One can also approximate a planning algorithm in the following way: For each initial condition and setting of the physical properties and forces, I consider the set of possible trajectories resulting from the agent taking any possible sequence of actions[13]. These trajectories are then scored under the different utility functions of the agents. An agent selects a trajectory according to its utility using a soft-max policy:

$$P(trajectory) \propto e^{-\beta \cdot U(trajectory))}, \tag{5.5}$$

where U is the utility function and β is a noise parameter [9, 186]. Utility functions will depend both on the state of the world and on the number of actions taken:

$$U(trajectory) = \sum_{i=1}^{N} d(state_i, state_{goal}) - k \cdot |F_i| \tag{5.6}$$

where $d(x, y)$ is a distance metric, F_i is the force the agent used to get to $state_i$ and k is a positive scaling parameter between actions and rewards. In the most general terms, such a utility function means that agents will prefer being 'close' to their goal state, that they will prefer to use as little force as possible to get there,

[13]For a single agent with A actions and N time-steps, there are $|A|^N$ trajectories to consider. A multi-agent problem would mean the product of actions, giving $(|A|x|A|)^N$ trajectories.

and that there is some trade off between the cost of using a force and the reward of being near or at the goal. The exact details of k and d could result in 'strong' and 'weak' agents, or 'motivated' and 'unmotivated' agents.

5.4.4 Resistive friction

So far, if an agent enacts a force and then stops, it will go on tumbling forever. There is something unappealing about the view that an agent can go tumbling along forever. As living creatures we constantly expend energy in an attempt to maintain balance. Thus, I introduce a *resistive friction* force f that acts against the motion of agents, such that the greater the force the agent produces, the greater the f working against it. Let us suppose for simplicity that $f(t) = -F_a(t-1)$. The dynamics of this friction are a discrete analog of two possible ways agents can move in the continuous case: They can either generate a very large but brief force at the onset and offset of motion ('impulses' in the physics engine terminology), or they might generate a sine-like force, leading to a 'biological' velocity profile of acceleration and deceleration [49].

5.4.5 An example scenario

Consider a simple launch-like event where an object of mass $m_1 = 1$ starts at position $x_1(0) = 0$ and with initial velocity $v_1(0) = +2\frac{c}{s}$. The object moves towards another object with mass $m_2 = 1$ at rest in position $x(0) = 6$. The two objects collide in between time steps 2 and 3, and the collision is resolved according to standard mechanics, thus the second object gains a velocity of $V(3) = +2\frac{c}{s}$ while the first object loses velocity and stays at rest.

Now instead, consider an agent at $x_1(0) = 0$ and an object at $x(0) = 6$. The

agent enacts a force $F_a = +2\frac{k*c}{s^2}$ at times $t = 0, t = 1$. This will bring the agent into contact with the object on the right. If it again enacts $+1\frac{c}{s^2}$ at time $t = 2$ it will cause the agent to have a velocity of $+2\frac{c}{s}$ at time $t = 3$, and thus between $t = 3$ and $t = 4$ we will need to solve a collision. There is some subtlety as to the question of 'what order' forces are resolved in during collision detection, but it can be resolved such that the agent will launch the object at $t = 4$.

While the two cases described will appear perceptually identical (assuming perception is limited to positions and velocities), the parsing of the scene is quite different in terms of forces and agenthood, depending on whether $entity_1$ is seen as a physical object with initial velocity, or as an agent enacting forces on itself.

5.4.6 Inference in Lineland

Given a generative model and some observations O of a trajectory T, we can use the standard Bayesian inversion to figure out the posterior probability over the hidden variables of interest, be they physical properties, goals, forces or something else. We will split the hidden variables into those that are about agents, ψ, and those that are about physics, θ. Agent variables include such things as goals and costs, physics variables include masses and non-agent forces.

The model's reasoning is captured by the following equation:

$$P(\psi, \theta|O) \propto P(O|\psi, \theta) \cdot P(\psi, \theta), \tag{5.7}$$

We cannot go directly from goals and physical parameters to observations, we

have to go through trajectories and a noisy observation function. Thus we have:

$$P(\psi, \theta|O) \propto \sum_{T} P(O|T) \cdot P(T|\psi, \theta), \tag{5.8}$$

where $P(T|\psi, \theta)$ is provided by the generative model, and $P(O|T) = \prod_i e^{-\delta(O_i, T_i)}$, δ being some distance metric, which here I will set to be the square distance.

If there are no agents involved in the scenario, the trajectory for a given setting of the physical parameters is deterministic[14]. However, if there is an agent involved, they could have made very poor decisions, and so we need to sum over all possible trajectories the agent could have generated through its action-sequence.

Because each agent could theoretically generate $|A|^N$ action-sequences, integration over all trajectories for all goals and initial conditions becomes difficult. But given a softmax decision policy with a reasonable β and a reasonable U, most action-trajectories for most goals will have such a minuscule probability of being chosen that we can consider only the top few action-sequences for any goal.

5.4.7 Inferring animacy in general

In Section 5.3 I noted that many of the cue-based accounts are concerned with finding classifiers/cues for a broad 'person/object' distinction. Such things can be quite simple, e.g. "If a shape has a face, it is animate".

While not denying that such cues exist, the general question of animacy might best be understood as a kind of hypothesis testing, summing over different sub-theories in the space of a generative like the one described above. While the generative model can be used to answer all sorts of specific questions about goals and

[14]Other proposals for physics-engines as the representation of intuitive physics considered noisy Newtonian-dynamics [12, 164].

properties, it can also aggregate over them. The model can infer that something is 'animate' rather than an 'object' without committing to any particular goal, if the sub-space of hypotheses in which the entity is an agent has greater probability than the sub-space of hypotheses in which it is an object.

So, the perception of animacy also includes a notion of an alternative, pure physics. In judging something to be animate, the model is effectively saying: "Regardless of a particular goal, the behavior of the entity is such that I am hard pressed to think of a purely physical explanation."

5.5 Discussion

The formal framework presented one particular model for reasoning about both physics and psychology. The model was applied to one particular domain, Lineland. But, we can use this model and framework to think more broadly about how intuitive physics and psychology interact. And they do interact, frequently. As adults we may not have to worry as much about inferring animacy as infants do, but we do think of agents as having mass, friction and other physical properties that constrain them, and help to make sense of their behavior. Maybe we use language faculties to quickly shuttle back and forth between our separate core domains [168], but pre-verbal infants can also make common use of these domains to 'make sense' of a scenario. How does this happen? How do these core systems interact?

One view already mentioned is the **multi-cue** view, the cues being organized either as a cascade or all cues firing at once. As I mentioned, there is an ongoing discussion whether cues alone could account for this kind of reasoning, but since cues are unlikely as a sole-explanation even within each domain [186, 12] it is doubtful

they can somehow combine together and create strength out of joint weakness[15].

"Two substances from the beginning...each follows only its own laws which it has received with its being, each agrees throughout with the other"
– Leibniz's Clock Analogy

Another view is more akin to **model selection**. Each core domain (objects and agents) is able to deal independently with the end product of 'objects moving over time'. Just as each core domain might have its own logical *not* operator rather than a cross-system shared operator [26], perhaps each domain has many other similar functions and operators. So, inference at a higher level might be selecting the core domain system to use, but once it is brought to bear on the task it performs all computation on its own. This implies both systems works in *parallel isolation*, with rapid switching of cognitive focus in cases like Lineland, as both systems compete to explain the stimuli.

A slightly different option is that of **independent but interacting** systems: both core domains are independent, but set up early on to interact and work jointly when explaining perceptual scenes. The formal model presented in this chapter approximates this view. Consider the concepts of *action* and *transition function.* Both are central to the domain of intuitive psychology, as modeled through inverse planning. Both are not necessarily tied to physics. An action for Mary can be "poison Sue's coffee", while the transition function might encode that this action leads to Sue's death. This allows rational inference over Mary's goals and motives, without any direct involvement from the physics system. But, actions can also be physical forces, and when they are they have a natural cost. The transition function could accept a ready-made 'step' function from the physics engine. So, while the concepts in each domain are distinct, they could be innately set up to interact with one another in specific ways.

הֵיְלכוּ שְׁנַיִם יַחְדָּו, בִּלְתִּי אִם-נוֹעָדוּ
-- עמוס ג' ג'

These different views have implications for future experiments, especially if sep-

[15] "The counsel of fools is all the more dangerous the more of them there are."
– Olaf the Peacock, *Laxdæla Saga*

arate brain regions can be associated with each core system. The *independent but interacting* view would suggest that when reasoning about scenes involving both physics and agents, these two core systems should somehow be able to refer to the same entity tokens. For example, the agent system would provide information regarding that entity's goal, while the physics system provides information regarding its mass, but both refer to the entity for a single computation of likely future paths. The *parallel isolation* view suggests both systems should be computing and considering future trajectories independently, each under its most likely candidate for explaining the world.

Beyond the question of how the core systems might interact, their similarities raise the speculative option that they start out as part of a single system.

5.5.1 Ur-system

The core systems of physics and psychology might be alike not just because of the evidence they are asked to explain, but because they are fundamentally the same system, or they branch out from one earlier system. This would be an "Ur-system" that is tasked with predicting and explaining the motion of entities. Perhaps the agent-system, with its core notions of utility and goals, is the primary system and objects are just a special kind of agent. Or perhaps agents are just objects, with very particular forces and potential functions.

Royal game of Ur

Objects from Agents

In this version of the argument, the basic units of the ur-system are utilities and agents. Every entity is assumed to have some goal, and objects are just very simple types of agents in terms of their utilities, actions and planning abilities. For example,

All things are full of gods
– Thales of Miletus

an object in a 'gravity well' is an agent that 'wants' to roll downhill, with the highest utility assigned to reaching the bottom of the well. This agent has some motive force, but is only able to plan locally. If it reaches a block in the middle of the hill, it cannot 'plan' beyond it and remains stuck in that position. The ur-system's biases continue to play out in the over-attribution of animacy, from ancient theories of the cosmos to the modern professor explaining the an electron as 'not wanting to share orbitals'[16]. The branching of the two domains (objects and agents) is then a process of restriction, where some agents are inferred to not have the full capacity of usual agents.

Agents from Objects

In this version, the basic units of the ur-system are forces and objects. Every entity is strictly moved by forces, whether internal or external. On this basis, some objects are categorized as belonging to a sub-class that has frequent internal forces. Alternatively, one can think of the basic units as being objects and potentials, where all objects simply follow greedy descent along a potential (forces and potentials are translatable). Most objects are guided by simple potential functions, but other objects have complicated potential functions. Such potential-field methods have been used in robotics to generate plans and navigate [104, 96]. The branching of the two domains (objects and agents) is then a process of generalization, where potential fields and movement options for the sub-category of motive objects are extended to more abstract spaces such as goal-space, configuration space and so on.

[16] "[W]hat makes planets go around the sun? At the time of Kepler some people answered this problem by saying that there were angels behind them beating their wings and pushing the planets around an orbit. As you will see, the answer is not very far from the truth. The only difference is that the angels sit in a different direction and their wings push inward." - Feynman, Lectures on Physics

Chapter 6

Afterword

"Fancy not its nature simple so" —
Lucretius discusses the mind, On
the Nature of Things

This work was concerned with formalizing intuitive theories: The space these theories live in, the cogs and wheels of our basic theories, and the induction mechanism that constructs new theories to understand the world. Some chapters focused more on the representation of knowledge, others on learning. But each chapter examined a mixture of these related ideas.

Back in the introduction I wanted to equip the reader with the concepts needed to make the trek through the thesis. I hope the reader also benefited from the journey itself, and saw some things along the way that are worth writing home about. We are at the end of the trek, but not near the summit: a full formal account of intuitive theories, core knowledge and theory learning. We don't know what the view from the top will be exactly, but looking back at the ground covered gives some sense of it.

On the issues of representation, there's good reason to think that core intuitive knowledge relates conceptual primitives through a generative framing, a kind of reverse-engineering of how the world works in a particular domain. There's also good reason to think that finding the right generative framing is hard, domain-specific work. Even in domains where a proposed frame exists (game engines for intuitive physics and a planning for agency), there's still a lot of conceptual, computational and experimental work left. For computational models, one generally-applicable advice is to consider for each knowledge domain how engineers had to tackle similar challenges in the 'forward' direction. As an example, software engineers need quick, cheap ways to simulate physics scenes. Their generative framework, along with tricks and hacks, make good candidates for mental frameworks of physics.

On the issue of learning, the stochastic search hypothesis put forward in Chapter 4 is almost certainly incomplete [151]. But the notion of taking the best current algorithms for searching structured, hierarchical spaces and examining them in the light of development and children's learning seems like the right way forward, regardless of the particular stochastic search proposal. Computational researchers and engineers have to come up with clever ways of searching conceptual spaces, and they likely come across the same sorts of difficulties and hacks that any rational agent would have to work through. While many of these researchers do not consider psychology, nearly every algorithm put forward in Machine Learning has been suggested by *someone* as 'the way the human mind does it'. These suggestions rarely make contact with child development, though, and could benefit from it.

So much for the general geography of formal intuitive theories, as seen from the current vantage point. Before closing, I want to point out some hazy things in the distance, avenues for future research within the general approach taken.

Intuitive psychology and intuitive sociology, agents and groups The building blocks of the core agency domain are goals, beliefs and contingent efficient actions. But there is much more to being a mental being. What about social circles, membership, in-group and out-group, hierarchies, dominance, imitation? These are distinct concepts that form a kind of 'intuitive sociology'. It has been suggested that this intuitive sociology is another core knowledge domain, as separate from agency as agency is from number [168, 131]. The general principles of this domain have not been worked out yet and remain the focus of much current research, but already we can ask some questions about the relationship of this domain to the domain of agency: Are agency and sociology distinct domains that operate in isolation? One domain? Two domains that communicate through some pre-established channels? What can a formal account tell us?

These questions are similar to the ones raised in Chapter 5 about the relationship between intuitive physics and intuitive psychology. Unlike that chapter, we don't yet have a concrete formal generative model of core social knowledge. One candidate is a graphical-network generator, of the sort that can generate cliques, chains, trees, hierarchies and so on to [88] reason about relations between entities. But even if we had such a worked out model, it still wouldn't be clear if it exists separately from a domain of agency, or physics for that matter. One possible way to link it with agency is through the prior on goals and beliefs. To see this, recall that the inference in the domain of agency was:

$$P(Goal, Belief|Action) \propto P(Action|Goal, Belief) \cdot P(Goal, Belief), \quad (6.1)$$

and that $P(Action|Goal, Belief)$ came from a planning principle. Where does $P(Goal, Belief)$ come from? Our probabilities over likely beliefs might come from a theory of perception, and our probabilities over likely selfish goals might come from some theory of hedonic sensation, but what about social goals? In Chapter 2 it was taken for granted that some agents help, some hinder and some are selfish. But surely we don't just have an arbitrary prior over these. We think people are generally likely to help, but also that they are likely to help people in their in-group and harm people in their out-group. The action of helping might be praiseworthy or blameworthy, depending on whether we helped a friend or enemy. Predicting, understanding, rewarding and punishing social behavior seems to require both a notion of social goals and a notion of group identity. So, the full treatment of reasoning about Heider and Simmel like stimuli will likely involve agency and physics, but also sociology.

Common sense in uncommon situations Our core intuitions and theories are not just about the world we find ourselves in. They are also about the worlds we can imagine, the close neighbors and far relations of the one we inhabit. As an example, imagine a wizard that can levitate a frog several feet off the ground, or conjure a frog into existence. Which of these feats is more impressive? Which takes more effort? We intuitively say the second as harder, but of course both are magic. Both are in some sense 'impossible'.

This is not a new claim, not in general nor in psychology. Walt Disney referred to instances of the 'plausible impossible', how we believe some imaginary things make more sense than other imaginary things. Some psychology researchers also see the imagination as a rational process that gets at the 'fault-lines of reality' [21].

Intuitive theories, especially those of psychology and physics, could be the underpinning of the plausibility of unreal worlds. So, notions of magic seem a particularly useful direction to explore as a way of getting at those intuitive theories. Creating a frog seems 'harder' than levitating it, perhaps because the first violates basic core object principles, while the second changes a property the object happens to have (location), furthest down on the hierarchy of properties described in Chapter 3. Our intuitive theories might also explain how we pick out the relevant properties within a violation, not just across violations. We know mass is the relevant property for levitation spells (it's harder to levitate a frog twice as massive, but not one twice as green), but in the case of an invisibility spell surface area might be more relevant than mass. Magic involving animate beings could tap into our intuitive psychology. Our sense of utility provides a natural metric for unnatural transformations of specific utilities (it's probably harder to magic someone into eating Brussels sprouts than it is to magic them into eating chocolate). Similar considerations might apply for belief and perception.

"The natural world has its laws, and no man must interfere with them...but they themselves may suggest laws of other kinds, and man may, if he pleases, invent a little world of his own, with its own laws" (George Macdonald, The Fantastic Imagination)

The previous paragraph was made of reasonable hunches, stitched together with pilot data, and it offers only a general roadmap for future research. It will take large-scale experiments, asking people to systematically rate the difficulty of various violations, to examine these ideas further.

The full understanding of the nature and origin of our intuition and common sense is still far off. But I am optimistic that the right formal tools for building this understanding are within reach.

Bibliography

[1] Edwin Abbott. *Flatland.* Broadview Press, 2009.

[2] Isabell EK Andersson and Sverker Runeson. Realism of confidence, modes of apprehension, and variable-use in visual discrimination of relative mass. *Ecological psychology*, 20(1):1–31, 2008.

[3] R. Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*, pages 47–83, 2002.

[4] Renée Baillargeon. How do infants learn about the physical world? *Current Directions in Psychological Science*, pages 133–140, 1994.

[5] Renée Baillargeon. Infants' physical world. *Current Directions in Psychological Science*, 13:89–94, 2004.

[6] Renée Baillargeon, Jie Li, Weiting Ng, and Sylvia Yuan. An account of infants physical reasoning. *Learning and the infant mind*, pages 66–116, 2009.

[7] Rene Baillargeon. Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, 3(1):2–13, 2008.

[8] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113:329–349, 2009.

[9] Chris L Baker, Rebecca R Saxe, and Joshua B Tenenbaum. Bayesian Theory of Mind. *SciencesNew York*, 1:1–10, 2011.

[10] CL Baker, ND Goodman, and JB Tenenbaum. Theory-based social goal inference. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1447–1452, 2008.

[11] H. Clark Barrett, Peter M. Todd, Geoffrey F. Miller, and Philip W. Blythe. Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, 26:313–331, 2005.

[12] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 110:18327–32, 2013.

[13] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 1st ed. 2006. corr. 2nd printing edition, October 2006.

[14] Ned Block. Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy*, 10:615–678, 1986.

[15] MGB Blum, MA Nunes, Dennis Prangle, and SA Sisson. A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

[16] Phililp. Blythe, Peter. Todd, and Geoffrey. Miller. how motion reveals intention. In *Social intelligence*, pages 257–285. 1999.

[17] Elizabeth Baraff Bonawitz, Tessa J P van Schijndel, Daniel Friel, and Laura Schulz. Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, 64:215–234, 2012.

[18] G F Bradshaw, P W Langley, and H A Simon. Studying scientific discovery by computer simulation. *Science (New York, N.Y.)*, 222:971–975, 1983.

[19] Cb Browne and Edward Powley. A survey of monte carlo tree search methods. *Intelligence and AI*, 4:1–49, 2012.

[20] J. S. Bruner, J. J. Goodnow, and G. A. Austin. *A study of thinking*. New York: Wiley, 1956.

[21] Ruth M. J. Byrne. The rational imagination and other possibilities, 2007.

[22] Susan Carey. *Conceptual change in childhood*. 1985.

[23] Susan Carey. *Conceptual change in childhood*. 1985.

[24] Susan Carey. Bootstrapping & the origin of concepts. *Daedalus*, 133(1):59–68, 2004.

[25] Susan Carey. Bootstrapping and the origin of concepts. *Daedalus*, 133(1):59–68, 2004.

[26] Susan Carey. *The Origin of Concepts.* Oxford University Press, 2009.

[27] Susan Carey and Elizabeth Spelke. Domain-specific knowledge and conceptual change. In *Mapping the Mind: Domain Specificity in Cognition and Culture*, pages 169–200. 1994.

[28] Erin Catto. Box2d: A 2d physics engine for games, 2013.

[29] Guillaume Chaslot, Sander Bakkes, Istvan Szita, and Pieter Spronck. Monte-Carlo Tree Search: A New Framework for Game AI. *AIIDE*, pages 216–217, 2008.

[30] Noam Chomsky. A Review of B.F. Skinner's Verbal Behavior. *Language*, 35:26–58, 1959.

[31] AM Collins and MR Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–247, 1969.

[32] Claire Cook, Noah D. Goodman, and Laura E. Schulz. Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 120:341–349, 2011.

[33] Gergely Csibra. Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, 107:705–717, 2008.

[34] Gergely Csibra, Szilvia Bíró, Orsolya Koós, and György Gergely. One-year-old infants use teleological representations of actions productively, 2003.

[35] D. C. Dennett. *The Intentional Stance.* MIT Press, Cambridge, MA, 1987.

[36] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[37] J. Feldman. An algebra of human concept learning. *Journal of Mathematical Psychology*, 50:339–368, 2006.

[38] Jerome A. Feldman, James G I Ps, James J. Horning, Stephen Reder, Jerome A. Feldman, James Gips, James J. Horning, and Stephen Reder. Grammatical complexity and inference. Technical report, Stanford University, 1969.

[39] H. Field. Logic, Meaning, and Conceptual Role. *Journal of Philosophy*, 74:379–409, 1977.

[40] Jzsef Fiser, Pietro Berkes, Gergo Orbn, and Mt Lengye. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119 – 130, 2010.

[41] J. Fodor and E. Lepore. Why meaning (probably) isn't conceptual role. *Mind & language*, 6(4):328–343, 1991.

[42] Jerry A. Fodor. *The language of thought.* Harvard University Press: Cambridge, MA., 1975.

[43] Jerry A. Fodor. On the impossibility of acquiring 'more powerful' structures. In *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky.* Harvard University Press, 1980.

[44] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[45] Kenneth D. Forbus. Qualitative Physics: Past, Present, and Future. In *Exploring Artificial Intelligence*, pages 239–296. 1988.

[46] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[47] Tao Gao, George E. Newman, and Brian J. Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59:154–179, 2009.

[48] Tao Gao and Brian J Scholl. Chasing vs. stalking: interrupting the perception of animacy. *Journal of experimental psychology. Human perception and performance*, 37:669–684, 2011.

[49] Gioele Gavazzi, Ambra Bisio, and Thierry Pozzo. Time perception of visual motion is tuned by the motor representation of human actions. *Scientific reports*, 3:1168, 2013.

[50] A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis Second Edition.PDF*, volume 1. 2004.

[51] G. Gergely, Z. Nádasdy, G. Csibra, and S. Biró. Taking the intentional stance at 12 months of age. *Cognition*, 56:165–193, 1995.

[52] György Gergely and Gergely Csibra. Teleological reasoning in infancy: The naïve theory of rational action, 2003.

[53] Sam Gershman, Ed Vul, and Joshua B. Tenenbaum. Perceptual multistability as Markov Chain Monte Carlo inference. *Advances in Neural Information Processing Systems*, 22:611 – 619, 2009.

[54] T. Gerstenberg, N. D Goodman, D. A. Lagnado, and J. B. Tenenbaum. Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[55] D L Gilden and D R Proffitt. Understanding collision dynamics. *Journal of experimental psychology. Human perception and performance*, 15:372–383, 1989.

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

[57] W.R. Gilks and DJ Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman & Hall/CRC, 1996.

[58] N. D. Goodman, V. K. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. *Uncertainty in Artificial Intelligence*, 2008.

[59] Noah D Goodman. Concepts in a probabilistic language of thought. *To appear in Concepts: New Directions*, 2014.

[60] Noah D. Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 2013.

[61] Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108—154, 2008.

[62] Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110, 2011.

[63] Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110–9, January 2011.

[64] A. Gopnik and A. N. Meltzoff. *Words, Thoughts, and Theories.* MIT Press, Cambridge, MA, 1997.

[65] A. Gopnik and D. Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 17(5):1205–1222, 2000.

[66] Alison Gopnik and Andrew N Meltzoff. Words, Thoughts, and Theories. *Mind: A Quarterly Review of Philosophy*, 108:0, 1999.

[67] Alison Gopnik and Laura Schulz. Mechanisms of theory formation in young children, 2004.

[68] Alison Gopnik and Henry M Wellman. Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085, 2012.

[69] Alison Gopnik and H.M. Wellman. The theory theory. In *Mapping the mind: Domain specificity in cognition and culture*, pages 257–293. 1994.

[70] Christian S. Gourieroux, Alain Monfort, and Eric Michel Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S):S85–118, 1993.

[71] Thomas L Griffiths, Elizabeth R Baraff, and Joshua B Tenenbaum. Using physical theories to infer hidden causal structure. In *proceedings of the 26th annual conference of the cognitive science society*, pages 500–505, 2004.

[72] Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14:357–364, 2010.

[73] Hyowon Gweon, Joshua B Tenenbaum, and Laura E Schulz. Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences of the United States of America*, 107:9066–9071, 2010.

[74] J. Kiley Hamlin. Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, 22:186–193, 2013.

[75] J. Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450:557–560, 2007.

[76] J.K. Hamlin and Karen Wynn. Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26(1):30–39, 2011.

[77] G. Harman. *Meaning and semantics.* New York University Press, New York, 1975.

[78] G. Harman. Conceptual role semantics. *Notre Dame Journal of Formal Logic*, 23:242–257, 1982.

[79] Fritz Heider and Marianne Simmel. An Experimental Study of Apparent Behavior. *The American journal of psychology*, 57:243–259, 1944.

[80] Susan Hespos, Gustaf Gredebäck, Claes von Hofsten, and Elizabeth S. Spelke. Occlusion is Hard: Comparing predictive reaching for visible and hidden objects in infants and adults. *Cognitive Science*, 33:1483–1502, 2009.

[81] Susan J. Hespos and Renée Baillargeon. Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition*, 107:304–316, 2008.

[82] David Hume. *A Treatise of Human Nature*, volume 26. 2000.

[83] Gavin Huntley-Fenner, Susan Carey, and Andrea Solimando. Objects are individuals but stuff doesn't count: Perceived rigidity and cohesiveness influence infants' representations of small groups of discrete entities. *Cognition*, 85:203–221, 2002.

[84] A. Jern and C. Kemp. Reasoning about social choices and social relationships. In P. Bello, M. Guarini, M. McShane, and B. Scassellati, editors, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

[85] Susan C Johnson, Virginia Slaughter, and Susan Carey. Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science*, 1:233–238, 1998.

[86] Y. Katz, N. D. Goodman, K. Kersting, C. Kemp, and J. B. Tenenbaum. Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, 2008.

[87] C. Kemp, N. D. Goodman, and J. B. Tenenbaum. Learning and using relational theories. *Advances in Neural Information Processing Systems*, 20:753–760, 2008.

[88] C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

[89] Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10:307–321, 2007.

[90] Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, 105:10687–10692, 2008.

[91] Charles Kemp and Joshua B. Tenenbaum. Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20 – 58, 2009.

[92] Charles Kemp, Joshua B. Tenenbaum, Sourabh Niyogi, and Thomas L. Griffiths. A probabilistic model of theory formation. *Cognition*, 114:165–196, 2010.

[93] J Kiley Hamlin, Tomer Ullman, Josh Tenenbaum, Noah Goodman, and Chris Baker. The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science*, 16(2):209–26, March 2013.

[94] J. Kiley Hamlin, Karen Wynn, and Paul Bloom. Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, 13:923–929, 2010.

[95] In Kyeong Kim and Elizabeth S Spelke. Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2):385, 1992.

[96] Jin-Oh Kim and Pradeep K Khosla. Real-time obstacle avoidance using harmonic potential functions. *Robotics and Automation, IEEE Transactions on*, 8(3):338–349, 1992.

[97] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[98] B. Klahr, D. & Macwhinney. Information Processing. In *Handbook of Child Psychology*, pages 631–678. 1998.

[99] K. P. Kording and D. M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

[100] V. Kuhlmeier, Karen Wynn, and Paul Bloom. Attribution of dispositional states by 12-month-olds. *Psychological Science*, 14(5):402–408, 2003.

[101] Thomas S Kuhn. History of Scientific Revolutions. In *History of Scientific Revolutions*, volume 3rd, page 226. 1996.

[102] Victor AF Lamme and Pieter R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579, 2000.

[103] R. Lattimore. *The Odyssey of Homer*. Harper Perennial Modern Classics, 1999.

[104] Steven M LaValle. Planning Algorithms. *Methods*, 2006:842, 2006.

[105] Steven M LaValle and James J. Kuffner. Rapidly-exploring random trees: Progress and prospects. In *4th Workshop on Algorithmic and Computational Robotics: New Directions*, pages 293–308, 2000.

[106] Sang Ah Lee and Elizabeth S Spelke. Two systems of spatial representation underlying navigation. *Experimental brain research*, 206(2):179–188, 2010.

[107] C. G. Lucas, A. Gopnik, and T. L. Griffiths. Learning the form of causal relationships using hierarchical bayesian models. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.

[108] C. G. Lucas and T. L. Griffiths. Developmental differences in learning the forms of causal relationships. *Cognitive Science*, 34:113–147, 2010.

[109] Christopher G Lucas, Sophie Bridgers, Thomas L Griffiths, and Alison Gopnik. When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2):284–299, 2014.

[110] Yuyan Luo, Lisa Kaufman, and Renée Baillargeon. Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology*, 58:441–486, 2009.

[111] David Marr. Vision. *book*, 1982.

[112] David Marr and Tomaso Poggio. From understanding computation to understanding neural circuitry. *AI Memo*, 357:1–22, 1976.

[113] Elena Mascalzoni, Lucia Regolin, and Giorgio Vallortigara. Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences of the United States of America*, 107:4483–4485, 2010.

[114] James L. McClelland. Parallel distributed processing: Implications for cognition and development. In *Parallel Distributed Processing: Implications for Psychology and Neurobiology*, pages 8–45. 1989.

[115] James L. McClelland, Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14:348–356, 2010.

[116] James L. Mcclelland and David E. Rumelhart, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models.* MIT Press, Cambridge, MA, 1986.

[117] JL McClelland. A connectionist perspective on knowledge and development. In *Developing cognitive competence: New approaches to process modeling*, pages 157—-204. 1995.

[118] Joseph McIntyre, M Zago, A Berthoz, and F Lacquaniti. Does the brain model newton's laws? *Nature neuroscience*, 4(7):693–694, 2001.

[119] A Michotte. *The perception of causality*, volume 6. 1963.

[120] R R Miller, R C Barnet, and N J Grahame. Assessment of the Rescorla-Wagner model. *Psychological bulletin*, 117:363–386, 1995.

[121] T.M. Mitchell. Generalization as search. *Artificial Intelligence*, (18):203–226, 1982.

[122] R. Moreno-Bote, D.C. Knill, and A. Pouget. Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108:12491–12496, 2011.

[123] G L Murphy and D L Medin. The role of theories in conceptual coherence. *Psychological review*, 92:289–316, 1985.

[124] Amy Needham and Renee Baillargeon. Intuitions about support in 4.5-month-old infants. *Cognition*, 47(2):121–148, 1993.

[125] A. Newell and H.A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19:113–126, 1976.

[126] Allen Newell. Physical symbol systems. *Cognitive Science*, 4:135–183, 1980.

[127] J. Pearl. Causality. *New York: Cambridge*, 2000.

[128] Ann T. Phillips and Henry M. Wellman. Infants' understanding of object-directed action. *Cognition*, 98:137–155, 2005.

[129] Jean Piaget and Bärbel Inhelder. *The psychology of the child.* Basic Books, 1969.

[130] Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123:199–217, 2012.

[131] Lindsey J Powell and Elizabeth S Spelke. Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, 110:E3965–72, 2013.

[132] David Premack and Ann James Premack. Infants Attribute Value to the Goal-Directed Actions of Self-propelled Objects, 1997.

[133] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming.* John Wiley and Sons, Inc. New York, NY, USA, 1994.

[134] Hilary Putnam. Brains and Behaviour. In *Readings in philosophy of psychology, Volume 1*, page 320. 1983.

[135] Lucia Regolin, Luca Tommasi, and Giorgio Vallortigara. Visual perception of biological motion in newly hatched chicks as revealed by an imprinting procedure. *Animal Cognition*, 3(1):53–60, 2000.

[136] R A Rescorla and A R Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II Current Research and Theory*, volume 21, pages 64–99. 1972.

[137] T. T. Rogers and J. L. McClelland. *Semantic cognition: A parallel distributed processing approach.* MIT Press, Cambridge, MA, 2004.

[138] David E Rumelhart and James L McClelland. On learning the past tenses of english verbs. 1985.

[139] David E Rumelhart, James L McClelland, and R J Williams. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. 1986.

[140] S Runeson, P Juslin, and H Olsson. Visual perception of dynamic properties: cue heuristics versus direct-perceptual competence. *Psychological review*, 107:525–555, 2000.

[141] S. Russell and P. Norvig. *Artificial Intelligence: a modern approach.* Prentice Hall, 3rd edition, 2009.

[142] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, Third edition.* 2014.

[143] MD Rutherford and Valerie A Kuhlmeier. *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention.* MIT Press, 2013.

[144] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120(2):411, 2013.

[145] Anne Schlottmann, Katy Cole, Rhianna Watts, and Marina White. Domain-specific perceptual causality in children depends on the spatio-temporal configuration, not motion onset. *Frontiers in Psychology*, 4, 2013.

[146] Anne Schlottmann, Elizabeth D. Ray, Anne Mitchell, and Nathalie Demetriou. Perceived physical and social causality in animated motions: Spontaneous reports and ratings. *Acta Psychologica*, 123:112–143, 2006.

[147] Brian J Scholl and Tao Gao. Percieving animacy and intentionality. In *Social perception: Detection and interpretation of animacy, agency, and intention.*, page 229. 2013.

[148] Brian J. Scholl and Patrice D. Tremoulet. Perceptual causality and animacy, 2000.

[149] J. Schultz, K. Friston, D. M. Wolpert, and C. D. Frith. Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. *Neuron*, 45:625–635, 2005.

[150] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[151] L. E. Schulz. *Finding new facts; thinking new thoughts*, volume 42. Elsevier, 2012.

[152] Laura Schulz. The origins of inquiry: Inductive inference and exploration in early childhood, 2012.

[153] Laura E Schulz and Elizabeth Baraff Bonawitz. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, 43:1045–1050, 2007.

[154] Laura E. Schulz, Noah D. Goodman, Joshua B. Tenenbaum, and Adrianna C. Jenkins. Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, 109:211–223, 2008.

[155] Laura E. Schulz, Alison Gopnik, and Clark Glymour. Preschool children learn about causal structure from conditional interventions. *Developmental Science*, 10:322–332, 2007.

[156] T. R. Shultz. *Cognitive Developmental Psychology.* MIT Press., Cambridge, MA, USA, 2003.

[157] Anna Shusterman, Sang Ah Lee, and Elizabeth S. Spelke. Cognitive effects of language on human navigation. *Cognition*, 120:186–201, 2011.

[158] R S Siegler and Z Chen. Developmental differences in rule learning: a microgenetic analysis. *Cognitive psychology*, 36(3):273–310, August 1998.

[159] R.S. Siegler and K. Crowley. The micro genetic method. *American Psychologist*, 46:606–620, 1991.

[160] Francesca Simion, Lara Bardi, Elena Mascalzoni, and Lucia Regolin. 3 from motion cues to social perception: Innate predispositions. *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*, page 37, 2013.

[161] Herbert A Simon. An information processing theory of intellectual development. *Monographs of the Society for Research in Child Development*, pages 150–161, 1962.

[162] B.F. Skinner. *Science and human behavior*, volume 80. 1953.

[163] E. Smith and D. Medin. *Categories and Concepts.* Cambridge, MA: Harvard University Press, 1981.

[164] Kevin A. Smith and Edward Vul. Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, 5:185–199, 2013.

[165] Beate Sodian, Deborah Zaitchik, and Susan Carey. Young Children's Differentiation of Hypothetical Beliefs from Evidence Development Young Children's Differentiation of Hypothetical Beliefs from Evidence. *Child Development*, 62:753–766, 2010.

[166] J. C. Spall. *Introduction to stochastic search and optimization: Estimation, simulation, and control.* John Wiley and Sons, 2003.

[167] Elizabeth S Spelke, Grant Gutheil, and Gretchen Van de Walle. The development of object perception. In *(1995). Visual cognition: An invitation to cognitive science, Vol. 2 (2nd ed.). An invitation to cognitive science*, pages 297–330. 1995.

[168] Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge, 2007.

[169] Elizabeth S. Spelke and S Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1990.

[170] Dan Sperber. In defense of massive modularity. *Language, brain and cognitive development: Essays in honor of Jacques Mehler*, 7:47–57, 2002.

[171] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, USA, second edition, 2001.

[172] Andreas Stuhlmüller and Noah D. Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 2013.

[173] Rashmi Sundareswara and Paul R. Schrater. Perceptual multistability predicted by search model for bayesian decisions. *Journal of Vision*, 8(5), 2008.

[174] Erno Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science (New York, N.Y.)*, 332:1054–1059, 2011.

[175] J. B. Tenenbaum and T. L. Griffiths. The rational basis of representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 1036–1041, 2001.

[176] J. B. Tenenbaum, T. L. Griffiths, and S. Niyogi. Intuitive theories as grammars for causal inference. In A. Gopnik and L. Schulz, editors, *Causal learning: Psychology, philosophy, and computation.* Oxford University Press, Oxford, 2007.

[177] Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10:309–318, 2006.

[178] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, 331:1279–1285, 2011.

[179] James T Todd and William H Warren. Visual perception of relative mass in dynamic events. *Perception*, 11(3):325–335, 1982.

[180] P. D. Tremoulet. *Animacy.* PhD thesis, Rutgers University, 2001.

[181] Patrice D. Tremoulet and Jacob Feldman. Perception of animacy from the motion of a single object. *Perception*, 29:943–951, 2000.

[182] Alan M Turing. Computing Machine and Intelligence. *MIND*, LIX:433–460, 1950.

[183] Claudia Uller and Shaun Nichols. Goal attribution in chimpanzees. *Cognition*, 76, 2000.

[184] Shimon Ullman. Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral cortex*, 5(1):1–11, 1995.

[185] Shimon Ullman, Daniel Harari, and Nimrod Dorfman. From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, 109(44):18215–18220, 2012.

[186] TD Ullman, CL Baker, and Owen Macindoe. Help or Hinder: Bayesian Models of Social Goal Inference. *NIPS*, pages 1–9, 2009.

[187] Tomer D Ullman, Noah D Goodman, and Joshua B Tenenbaum. Theory learning as stochastic search in the language of the thought. *Cognitive Development*, 2012.

[188] R F Wang, L Hermer, and E S Spelke. Mechanisms of reorientation and object localization by children: a comparison with rats. *Behavioral neuroscience*, 113:475–485, 1999.

[189] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[190] Henry M. Wellman, Fuxi Fang, and Candida C. Peterson. Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives. *Child Development*, 82:780–792, 2011.

[191] Henry M Wellman and Susan A Gelman. Cognitive development: Foundational theories of core domains. *Annual review of psychology*, 43(1):337–375, 1992.

[192] Henry M. Wellman and Jacqueline D. Woolley. From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3):245–275, 1990.

[193] HM Wellman and SA Gelman. Knowledge acquisition in foundational domains. In *The Handbook of Child Psychology*, pages 523–573. 1998.

[194] A L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69:1–34, 1998.

[195] Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological review*, 114:245–272, 2007.

[196] Wako Yoshida, Ray J. Dolan, and Karl J. Friston. Game theory of mind. *PLoS Computational Biology*, 4(12):1–14, 2008.

[197] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10:301–308, 2006.

[198] Jeffrey M. Zacks. Using movement and intentions to understand simple events. *Cognitive Science*, 28:979–1008, 2004.