Abstract
The Social Representation of the Physical World
Michael Anthony Lopez-Brau
2023

People have a remarkable capacity to reason about others' mental states—their beliefs, desires, and goals—from simply observing their behavior. However, this capacity extends beyond scenarios where other people are directly observable: A coat hanging over a chair, traffic cones blocking a lane on the road, and belt barriers at the airport all elicit rich social inferences about an agent, despite appearing as nothing more than physical objects. While previous research has documented our capacity to extract social information from physical objects, little research has investigated the computations and representations that underlie this capacity. In this thesis, I propose that three interconnected inferences support our capacity to build and reason about social representations from physical objects. I elucidate the mechanisms behind these inferences and show how they can be explained through a combination of our specialized cognitive capacities to process physical and social information.

The Social Representation of the Physical World

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Michael Lopez-Brau

Dissertation Director: Julian Jara-Ettinger

May 2023

To my father, who always supported me in the toughest of times.

Acknowledgment

I must begin these acknowledgements by thanking my parents. Neither of them had an ‘easy life’ growing up, but they persevered and carved their place in the world. Throughout my life, they’ve always ensured that my necessities were always met. I’ll never forget the sacrifices they made to shape the person who I am today.

I owe a massive ‘thank you’ to my advisor: Julian. We both embarked on a new journey around the same time: you as an advisor and me as your student. I learned so much from you over the years and I know that I’ll greatly miss talking about research together. I can’t begin to express my gratitude for your patience, empathy, and support during my time at Yale.

I’d also like to thank Brian. Even though we didn’t get the chance to work together as much as I wanted to, I still greatly appreciated our collaboration. From you I learned a great deal about science, best research practices, and the inferiority of Windows machines. Your lab meetings were always intellectually stimulating and enjoyable. Thank you for letting me be a part of it all.

More broadly, thank you to the supportive faculty at Yale that have served on my many committees—Ilker, Laurie, Maria, and Sam—as well as to the rest of the faculty that supported me during my time here. I’ve also been fortunate to have been surrounded by a supportive student cohort. Starting graduate school is challenging in many ways, but one of the hardest is being thrown into an unfamiliar environment with unfamiliar people. I’d like to thank the friends I made along the way for making the journey so much better.

First, thank you to the old and current members of Computational Social Cognition Lab. Rosie, I’ll never forget the many ‘pranks’ you executed, especially the day I walked into lab and discovered that you’d hidden cutouts of my face around the room. You probably played the biggest role in helping me settle in, inviting me to all sorts of brunches/lunches/dinners at your place, outdoor events, and department events. I truly thank you for that. Amanda, thank you for being someone I could ‘keep it real’ with and talk to about anything (i.e., the “hot goss”). I’ll never forget our once-in-a-lifetime research trip to Bolivia. I’m thankful that you decided to join the lab again for graduate school. Marlene, thank you for being such a warm and kind friend. I’ll miss our random conversations in the lab, your get-togethers, and getting to pet Olive. Flora, thank you for the many long conversations and emotional support. Madison and Colin, thank you for being such great lab managers, dealing

with matters ranging from using the p-card and handling reimbursements to debating where the trash can and recycling bin should go. I can't forget about our wonderful post-docs—Isaac, Mika, and Daniel—for sharing their graduate school experiences and job market advice. Finally, thank you to Paula and Georgina, two of the lab's collaborators and friends.

Next, thank you to my developmental cohort: Alyssa, Emily, Emory, Gracie, and Kate. Alyssa, thank you for always encouraging me to go on a run (even though I would almost always decline). Emily (and Daniel!), thank you for all of the get-togethers you hosted. Emory, Gracie, and Kate, thank you for the conversations during the many food and drink outings we shared. A little more briefly, I'd like to also thank my cognitive cohort: Brynn and Joan. Thank you to Vlad and Ryan for being great friends who were always down to hang out (and also hosted amazing parties). A big thank you to the newer members of the department—Mario, Zach, Mandy, Tristan (and Paul!), Kat, Aalap, Kim, and Merve. I also can't forget to mention the previous members of the department who played a critical role. Stefan, I only wish we'd overlapped for longer, but I'm thankful that we at least got to know each other before you left. Sami, right up there with Rosie, you were one of the people that helped me settle in the most. I enjoyed our many trips to Bruegger's bagles, and to this day I have yet to meet someone that has as strange of an order as you. Cameron, I had a lot of fun talking science, debating 'random' topics in fancy attire, and playing 'Smash'.

Finally, I'd like to thank people outside of Yale. Breanna, since you've entered my life, you've helped me find peace in the craziness that is life. I look forward to spending many, many more years together. Carlos, you've always been an incredible friend, ready to hang out at a moment's notice during my often-sudden trips back home. Nate, you're hands-down one of the best friends anyone could have, always willing to listen to my ramblings and offer your perspective and emotional and physical support. In the rush to submit my dissertation, there are more relationships and experiences that I didn't recount here, but I am thankful to everyone that positively shaped my graduate school experience.

Contents

1 Introduction 1

1.1 Intuitive theories . . . 3

1.1.1 Intuitive psychology . . . 4

1.1.2 Intuitive physics . . . 5

1.2 Detecting social information from objects . . . 6

1.2.1 Cues to agency . . . 6

1.2.2 Agency and order . . . 7

1.2.3 “Human error” hypothesis . . . 8

1.3 Detecting social information through physical reasoning . . . 9

1.4 Inferring mental states from objects . . . 9

1.5 Thesis overview . . . 12

1.5.1 Chapter 2 . . . 12

1.5.2 Chapter 3 . . . 13

1.5.3 Chapter 4 . . . 13

1.5.4 Chapter 5 . . . 14

1.5.5 Chapter 6 . . . 15

2 How do we detect the presence of agents? 16

2.1 Introduction . . . 16

2.2 Experiment 1a . . . 18

2.2.1 Participants . . . 18

2.2.2 Stimuli . . . 18

2.2.3 Procedure . . . 19

2.2.4 Results . . . 19

2.3 Experiment 1b . . . 20

2.3.1 Participants . . . 20

2.3.2 Stimuli . . . 21

2.3.3 Procedure . . . 21

2.3.4 Results . . . 21

2.4 Discussion . . . 22

3 Detecting social information through physical reasoning **24**
3.1 Introduction 1
3.2 Computational Framework 2
3.3 Behavioral Experiment 5
3.3.1 Participants 5
3.3.2 Stimuli 5
3.3.3 Procedure 6
3.3.4 Results 6
3.4 Discussion 7

4 How do we reconstruct the past actions of agents? **10**
4.1 Introduction 1
4.1.1 Connection to related proposals in social psychology 2
4.1.2 The current work 3
4.2 Computational Framework 3
4.2.1 Implementation Details 6
4.3 Experiment 1a 6
4.3.1 Participants 6
4.3.2 Stimuli 6
4.3.3 Procedure 7
4.3.4 Results 8
4.4 Experiment 1b 10
4.4.1 Participants 10
4.4.2 Stimuli 10
4.4.3 Procedure 11
4.4.4 Model Predictions 12
4.4.5 Results 12
4.5 Experiment 2 12
4.5.1 Participants 13
4.5.2 Stimuli 13
4.5.3 Procedure 13
4.5.4 Model Predictions 14
4.5.5 Results 14
4.6 Do explicit event reconstructions in Experiment 2 predict inferences from Experiment 1? 16
4.7 Experiment 3 17
4.7.1 Participants 17
4.7.2 Stimuli 17
4.7.3 Procedure 18
4.7.4 Model Predictions 19
4.7.5 Results 19
4.8 Discussion 21
4.8.1 What cognitive capacities are required for event reconstruction? 22

4.8.2 Study limitations 24
4.8.3 Implications and conclusions 25

5 How do we infer the communicative meaning of objects? 27
5.1 Introduction 1
5.1.1 Paper overview 4
5.2 Computational Framework 4
5.2.1 Non-mentalistic decider 6
5.2.2 Simple enforcer 6
5.2.3 Mentalistic decider 7
5.2.4 Complex enforcer 8
5.2.5 Model implementation details 9
5.3 Model Analysis 9
5.4 Experiment 1: Quantitative model evaluation 11
5.4.1 Participants 12
5.4.2 Stimuli 12
5.4.3 Procedure 13
5.4.4 Model Predictions 14
5.4.5 Results 15
5.5 Overview of Experiments 2-4 18
5.6 Experiment 2: Are objects that impose a cost more likely to be communicative? 19
5.6.1 Participants 19
5.6.2 Stimuli 19
5.6.3 Procedure 20
5.6.4 Results 20
5.7 Experiment 3a: Are low-cost objects interpreted as communicative blockers? 21
5.7.1 Participants 21
5.7.2 Stimuli 21
5.7.3 Procedure 21
5.7.4 Results 22
5.8 Experiment 3b: Replication with the Tsimane' 22
5.8.1 Participants 23
5.8.2 Stimuli 23
5.8.3 Procedure 23
5.8.4 Results 23
5.8.5 Discussion 23
5.9 Experiment 3c: Inferences from conventional communicative blockers 24
5.9.1 Participants 25
5.9.2 Stimuli 25
5.9.3 Procedure 25
5.9.4 Results 25

5.10 Experiment 4: Conventionalizing object meanings 26
5.10.1 Participants 27
5.10.2 Stimuli 27
5.10.3 Procedure 27
5.10.4 Results 28
5.11 Experiments 2-4 Discussion 28
5.12 General Discussion 29
5.12.1 Model assumptions and study limitations 29
5.12.2 Open questions 32
5.13 Concluding Remarks 33
5.14 Acknowledgments 33

6 Discussion 34
6.1 Chapter 2-5 Review 35
6.2 Open questions 36
6.2.1 How do intuitive psychology and intuitive physics combine? . 36
6.2.2 Is this capacity uniquely human? 36
6.2.3 Context in culture 38
6.2.4 Parsimony in explanation: heuristics versus Bayesian inference 38
6.2.5 Efficient computation and amortized inference 39
6.2.6 When and how do agents come into the picture? 40
6.2.7 Inferring mental states beyond communicative intent 41
6.2.8 Detecting intentional action 41
6.3 Conclusion 42

Chapter 1

Introduction

Imagine hiking along a trail in a mountainous forest. Along your trek, the trail becomes progressively less marked, making it difficult to determine where the path continues. As you begin to question whether you went the wrong way, you notice a small stack of rocks, about three feet in height (Figure 1.1). Intuitively, this simple arrangement of objects immediately grabs your attention, not because of its precarious assembly, but because it reveals that someone was here and stacked the rocks (as rocks cannot stack naturally, bar an appropriate causal process, such as residing at the base of a cliff). Beyond being able to detect the involvement of an agent, you can also infer how the agent assembled it (e.g., by stacking one rock on top of the other, favoring larger and flatter rocks at the base), how long it took to build, and how much effort was involved. Finally, you can even infer that the agent did so with the purpose of communicating their mental states (e.g., to communicate where the trail continues).

Similar experiences are ubiquitous in our everyday life: rope surrounding a patch of grass on a golf course tells us not to walk through, traffic cones on the road guide where we drive our cars, and stanchions at a movie theater mean we should form a line (see Figure 1.2 for more examples). While these examples are widespread, they pose a challenge to existing theories of mental representations in cognitive science, which emphasize that social and physical reasoning are separate cognitive systems. In infants, these systems are isolated, encapsulated, and form a subset of our *core knowledge* (Spelke, 2003; Spelke and Kinzler, 2007), and, in adults, they are instantiated by different neural circuity (Saxe and Kanwisher, 2003; Fischer et al., 2016). Moreover, previous work suggests that social reasoning might be supported by bottom-up processes that directly detect agents in our environment, such as through a sensitivity for faces or eyes (Johnson et al., 1991; Simion et al., 2001; Colombatto and Scholl, 2022). However, as the examples in Figure 1.2 show, people can also spontaneously detect that physical objects contain social information, enabling further social reasoning about the agent that was involved.

Consistent with this, past research has shown that people have a rich understanding of what physical environments reveal about others (Gosling et al., 2002; Hurwitz and Schachner, 2020). For instance, people can infer others' actions and goals from

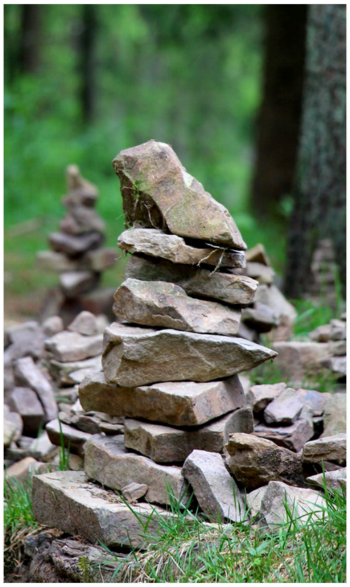

Figure 1.1: A stack of rocks also known as a cairn. These structures have historically served as landmarks for a variety of purposes (e.g., burials, hunting, storage). Today they are typically used as trail markers.

indirect physical evidence of their presence (Lopez-Brau et al., 2022) and estimate the effort involved in moving and manipulating objects (Yildirim et al., 2019). These capacities also emerge early in development, with children drawing surprisingly rich inferences from physical evidence, ranging from inferences about what actions an agent took (Jacobs et al., 2021) and what they knew (Pelz et al., 2020) to inferences about even richer social information, such as whether two people transmitted ideas (Pesowski et al., 2020) and have shared interests (Pesowski et al., 2021). While this large body of work has documented our capacity to infer and reason about social information from physical objects, little work has explored what computations and representations support this capacity.

In this dissertation, I propose that our capacity to build and reason about the social representations of physical objects is supported by three interconnected inferences. First, from these objects we are able to detect that they contain social information—it is obvious that an agent was involved in arranging them and not a natural force, like the wind. This is critical for knowing when to reason about another agent, as opposed to doing so for every object in our visual experience. I review this capacity in Chapters 2 and 3. Second, we can mentally reconstruct how the agent manipulated the scene. This can be as simple as inferring where an agent was walking to or from based on mud tracks that they left behind, but can also be as sophisticated as estimating the time and effort behind a marble sculpture. I review this capacity in Chapter 4. Finally, we can reason over the agent's reconstructed actions to infer their mental states. I review this capacity in Chapter 5. I propose that, together, these three inferences enable people to embed and extract rich social information

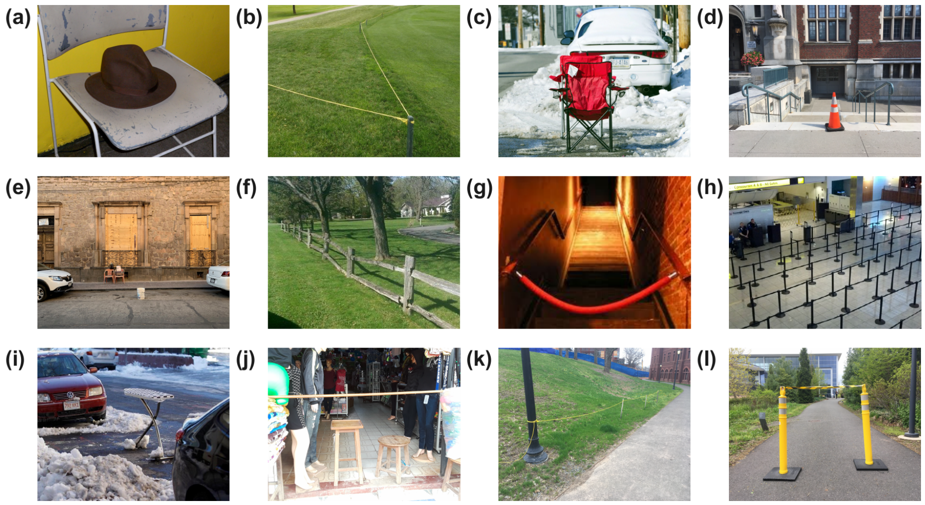

Figure 1.2: Real-world examples of people using objects to relay social information. (a) A hat on a chair indicating that someone intends to return. (b) Rope a few inches above the grass so that people know not to walk through. (c) Chairs along the side of the street in South Boston to reveal someone shoveled and claimed this parking spot. (d) A traffic cone in front of some stairs signaling limited access. (e) A bucket along the side of the street in central Mexico indicating that the parking spot is reserved. (f) An easy-to-cross fence marking a property limit. (g) A stanchion across a stairwell revealing access may be restricted to certain individuals. (h) Belt barriers at the airport telling passengers that they should form a line (and where). (i) An ironing board along the side of the street indicating that the parking spot is taken. (j) A wooden pole and two small benches in a store in Bolivia indicating that the owner is not available. (k) A small rope along a sidewalk asking people not to walk near a construction site. (l) A pair of traffic posts preventing people from using this walkway.

from the physical world in a way that is uniquely human. In the remainder of this introduction, I present the theoretical foundation to my proposal (see Table 1.1 for key concepts), followed by a brief overview of each component inference, and close with a brief summary of each chapter.

1.1 Intuitive theories

A mainstream idea in cognitive science is that humans are born with *core knowledge*, a foundational understanding over several core domains (Spelke, 2003; Spelke and Kinzler, 2007; Lake et al., 2017). These core domains include number (numerical and set operations), space (environment geometry and navigation), psychology (agents and their actions), physics (objects and their mechanics), and, more recently, sociology (social groups and their interactions). Critically, each system is encapsulated—the internal workings of each system are largely inaccessible to other representations and computations—and isolated—representations constructed by each system do not readily combine with one another. These core knowledge systems further support our

Bayes' theorem: Describes the probability of an event A given prior knowledge of another event B ($p(A|B)$; also known as the *posterior* or posterior probability). Mathematically, the posterior is proportional to the likelihood of the second event given the first event ($p(B|A)$) multiplied by the probability of the first event ($p(A)$).

Inverse planning: In forward planning, agents choose actions according to their goals, beliefs, and desires. In inverse planning, agents work backwards, taking observable actions and simulating possible goals, beliefs, and desires to find the set that best explains the observed data.

Generative models: A model that represents the probabilistic distribution of some observable data. In this thesis, this is synonymous with the likelihood term in Bayes' theorem.

Intuitive theories: Abstract causal mental models that provide understanding for phenomena across various domains, such as psychology, physics, and biology.

Intuitive psychology: The foundational capacity to reason about an agent's behavior in terms of their mental states.

Intuitive physics: The foundational capacity to reason about objects and their physical interactions.

Physics engine: Computer engines that approximately simulate Newtonian physics in video games and graphics.

Table 1.1

intuitive theories, abstract causal models that guide the interpretation, explanation, and prediction of relevant phenomena (see Table 1.1). Intuitive theories are similar to scientific theories, consisting of an ontology of concepts and a system by which they are related (Wellman, 1992; Carey, 2009; Gerstenberg and Tenenbaum, 2017). In particular, my proposal focuses on the combination of two intuitive theories—intuitive psychology and intuitive physics—and how this combination supports our capacity to reason about social information from physical objects. That two intuitive theories may work together has interesting implications on our understanding of core knowledge. I discuss this in the Chapter 6. Before introducing how this combination might work, I briefly review the empirical and theoretical evidence for each intuitive theory, including recent attempts to formalize them as computational models.

1.1.1 Intuitive psychology

People have an intuitive understanding of others. Early in development, we expect others to act efficiently towards their goals, guided by a principle of rational action (Gergely et al., 1995; Gergely and Csibra, 2003). 15-month-olds start to understand more complex representations, such as false beliefs (Onishi and Baillargeon, 2005; Scott and Baillargeon, 2017), and 18- and 20-month-olds understand that other people have desires, which form the motivation for their actions, that differ from their own (Repacholi and Gopnik, 1997; Kushnir et al., 2010). Furthermore, we can invert this intuitive causal model of others to explain their behavior in terms of their mental states: their goals, beliefs, and desires (Wellman, 2014; Gopnik et al., 1997). Previous work suggests that our intuitive psychology is supported by both bottom-up processes, such as through visual cues to agency (New et al., 2007; Johnson et al., 1991; Troje, 2013; Johansson, 1973, 1976; Colombatto and Scholl, 2022; Scholl and Gao, 2013), and top-down processes, such as through the use of our mental causal models of others, also known as our *Theory of Mind*.

Recent research has attempted to formalized intuitive psychology as *Bayesian inverse planning* (Baker et al., 2009; Baker et al., 2017; see Table 1.1). Assuming that agents act efficiently towards their goals, we can form reasonable expectations for what actions agents will take given their beliefs and desires. For instance, consider several food trucks parked on a street. Knowing that someone's favorite food is Lebanese makes their next sequence of actions obvious: they will take a path towards the Lebanese food truck. This action sequence is derived by using our causal model, here also known as forward planning. However, in many situations, the mental states are not known and all we have access to is the sequence of actions that an agent took. To infer someone's mental states from their behavior, we must invert our causal model, here also known as inverse planning. In more recent work, it has been shown that people not only infer others' mental states from their actions, but also make graded inferences about the costs and rewards that were involved, known as a *naïve utility calculus* (Jara-Ettinger et al., 2016, 2020a). Together, these computational frameworks form the foundation for social reasoning within the models I propose.

1.1.2 Intuitive physics

Similar to intuitive psychology, people also have an intuitive understanding of physical objects and their mechanics. Intuitive physics is also available early in development (Téglás et al., 2011; Spelke and Kinzler, 2007), with infants demonstrating an understanding of the spatio-temporal principles of cohesion (objects move as connected and bounded wholes), continuity (objects move on connected, unobstructed paths), and contact (objects do not interact at a distance; Aguiar and Baillargeon, 1999; Leslie and Keeble, 1987; Spelke, 1990). While our intuitive understanding of physics becomes more sophisticated in development, it remains only a rough approximation to classical or Newtonian physics (Levillain and Bonatti, 2011).

Recent work has attempted to model intuitive physics using a *physics engine* (Battaglia et al., 2013). Physics engines are computer engines that simulate Newtonian physics in video games and graphics using approximate, probabilistic simulations. In this study, people were presented with images of block towers and were tasked with judging whether it was stable and, if it was not stable, where it would fall. Their proposed intuitive physics engine framework yielded strong quantitative fits with participant judgments—even when the mass, shape, and force applied to the block towers was varied—revealing how this framework not only affords us the flexibility to make noisy inferences about physical parameters (e.g., what is the speed of a car?), but also about entire scenes (e.g., will two cars collide and would a collision cause collateral damage?). This computational framework supports some of the physical reasoning within the models I propose.

1.2 Detecting social information from objects

These intuitive theories support reasoning about social and physical entities, but most of the empirical evidence has focused on how they do so individually rather than jointly. Before we can reason about the mind behind an object, however, we must know that there is a mind to reason about in the first place. In the examples of Figure 1.2, perhaps the most obvious inference we can draw is that an agent was involved in arranging the scene. Why is this so obvious? One possibility is that there exist low-level visual cues in the objects that agents interact with that reveal their involvement. This idea presumes that our visual system functions as a filter for these low-level visual cues: the *traces of agency*. By contrast, if these traces did not receive specialized visual processing, we would be overwhelmed by the sheer magnitude of objects in our visual environment. A common way to prioritize a stimulus is through the selective operation known as attention, which acts a filter to both what we perceive and what we think about (Chun, 2011; Treisman, 2006). The idea of specialized visual processing for certain types of stimuli is not a new idea—in fact, decades of work in vision science has shown that our visual system prioritizes the most social stimuli around: people (New et al., 2007; Johnson et al., 1991; Troje, 2013; Johansson, 1973, 1976; Colombatto and Scholl, 2022; Scholl and Gao, 2013). If agents receive specialized visual processing, might some form of this specialization also extend to the traces of agency and, if so, what are the cues?

1.2.1 Cues to agency

Perhaps the most well-studied cue to agency is faces. Our brain devotes an incredible amount of resources to detecting faces, up to the point of having an entire brain region specialized for face recognition: the fusiform face area (Kanwisher et al., 1997). Our proficiency at classifying faces can be partly attributed to the developmental trajectory of this capacity; just 30 minutes after birth, infants already show preferential tracking of face-like patterns (consisting of an oval with three dots, two for the eyes and one for the mouth), following a moving face-like pattern farther than other moving patterns of similar complexity (Johnson et al., 1991). While more recent work has challenged the three-dot arrangement, it continues to support that infants generally prefer “top-heavy” geometries (Simion et al., 2001).

The way that people move is also a cue to what and who they are, known as biological motion (Troje, 2013). Previous work revealed that from a moving arrangement of fewer than 10 lights, people could perceive a fully coherent shape of a human figure (Johansson, 1973, 1976). Multiple brain regions are activated during biological motion perception, with the primary region being the superior temporal sulcus (Bonda et al., 1996; Grossman et al., 2000; Grossman and Blake, 2002; Grossman et al., 2004; Peuskens et al., 2005).

These cues both map onto aspects of the human body (i.e., its shape), however self-propulsion is one cue to agency that is completely independent from it, and one that infants and non-human primates also rely on (Premack, 1990; Hauser, 1998; Di Giorgio et al., 2017). In the seminal study by Heider and Simmel (1944), participants were presented with a video of several moving shapes, moving in a particular pattern. Despite having no other information available, people readily interpret these shapes as agents (e.g., interpreting one as an aggressor), and attribute sophisticated mental states to them (e.g., dominance). In another study, Gao et al. (2010) found that participants irresistibly perceived an array of triangles as intentional and goal-directed when oriented towards them, impairing their performance in interactive tasks. A related study further showed that even when the triangles were irrelevant to the task, participants' visuomotor behavior was still impaired (van Buren et al., 2016). From this body of work, it is clear that perceiving agency does not require seeing a living being—or even a proxy to one, as in Troje (2013).

These findings highlight the many ways in which the mind automatically and irresistibly detects agency (and even intentionality) from visual stimuli consisting of nothing more than simple shapes. All of this work, however, is about how we directly detect agents. In the examples presented earlier (Figure 1.2, we instead indirectly detect that an agent was present, because there is no agent (of any kind) to observe. This raises the question of whether the capacity to directly detect agents extends to the ability to infer their involvement from the way they manipulate the environment.

1.2.2 Agency and order

People manipulate their environment in a variety of ways, but generally have a tendency to increase the order within them. We see examples of this in the consistent timing of a traffic light, changing from red to green (temporal order), the neat spatial arrangement of kitchen tiles inside of a house (spatial order), or in mathematical constructs, like an infinite series (conceptual order). What is common across these distinct forms of order is that they all contain regularity. How might order from regularity reveal the involvement of agents?

Previous work has shown that from a very young age, children understand the link between agency and order (Newman et al., 2010; Keil and Newman, 2015a). In Newman et al. (2010), 12-month-old infants were presented with groups of red and blue blocks and a ball that either had animate features (e.g., eyes) or did not. Infants would witness either the animate or inanimate ball arrange the group of blocks from an disordered state (e.g., with the blocks in a haphazard pile) to a ordered one (e.g., with the blocks arranged into two neat columns, separated by color). Infants looked

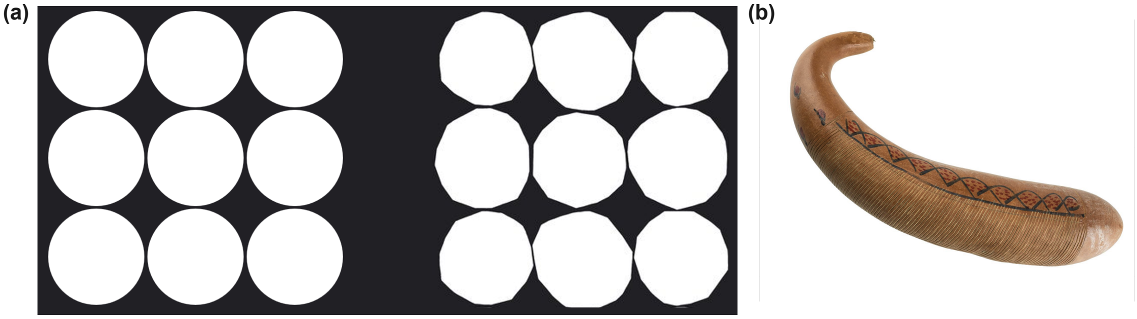

Figure 1.3: (a) Example of the contrast between perfect order and order with slight errors. The grid of circles on the left appear so perfectly made, an agent would require a stencil to create. The grid of circles on the right, however, are just the kind of thing that an agent with a pair of scissors would produce. (b) A güiro, a Latin-American instrument often handmade from either wood or fiberglass, is one of the many real-world examples of subtle imperfections in art and music.

longer when the inanimate ball was responsible for the state change, but would not exhibit preferential looking when the animate ball was responsible for the same action. In related work, 9-month-old infants associated a regular color sequence (e.g., yellow, yellow, red, yellow, yellow, red) as being the product of a human hand instead of a mechanical one (Ma and Xu, 2013). Taken together, these findings reveal that even infants expect animate entities to be "creators of order" and inanimate entities to be "creators of disorder".

1.2.3 "Human error" hypothesis

This suggests that a potential cue for the traces of agency is whether or not an object or environment contains order. However, maximal order sometimes seems non-agentive, instead likely to be the work of a machine or even a supernatural entity (Aquinas, 1485; Dawkins et al., 1996). Furthermore, this idea neglects a key signature of human involvement: the errors introduced by their actions. That is, people cannot perfectly manipulate their environment, so whenever they act towards their goals, they leave behind errors.

Suppose that Figure 1.3a is a black piece of construction paper with holes in it. The holes on the left seem unlikely to have been made by a person (without a specialized tool). The holes on the right—with many slight imperfections—seem like just the kind of outcome that we would expect from a person using a pair of scissors. These kinds of errors make their appearance in other places (e.g., Figure 1.3b). In folk art and aesthetics, handmade crafts are seen as more valuable, a homage to pre-industrial societies where knowledge and skills were personal (Oring, 1986). Even music is rich with subtle imperfections, such that electronic artists add noise to their tracks to simulate analog instrumentation. The first goal of this thesis is to investigate if objects containing order—with small imperfections—are prioritized by our visual system (Chapter 2).

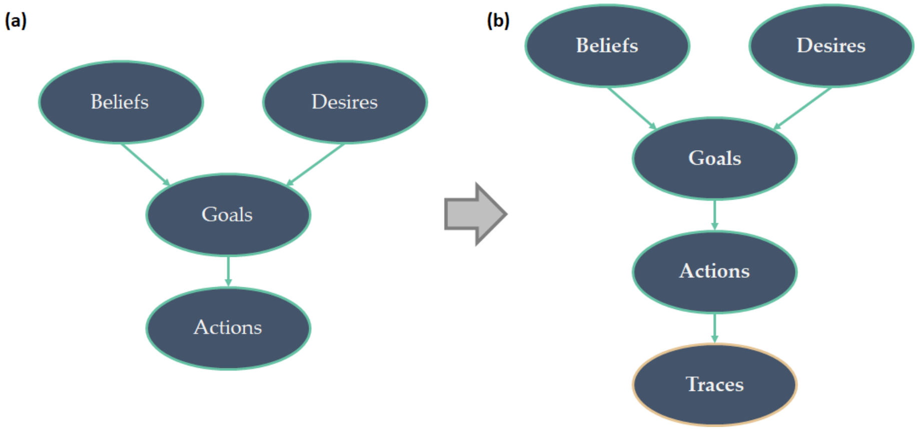

Figure 1.4: Conceptual representation of our model extension. (a) Traditional belief-desire psychology framework (Fodor, 1992; Wellman and Bartsch, 1988). Agents form goals on the basis of their beliefs and desires, and these goals serve as motivation for their actions. (b) When actions are unobservable, agents must rely on the physical traces that agents leave behind in order to reconstruct their actions. With these reconstructed actions, agents can then apply Theory of Mind over them to infer goals, beliefs, and desires.

1.3 Detecting social information through physical r-easoning

Superficial visual properties might not be the only way to detect whether an object contains social information. Some objects reveal that an agent was involved by considering the alternative: Could some natural process do this? That is, are people first analyzing their environment through physical reasoning and engaging in social reasoning whenever they detect features of the environment that cannot be explained by an appropriate natural process?

Consistent with this hypothesis, previous research suggests that violations of physics trigger an expectation of agency. For instance, after seeing an inert object fly into a scene, infants expect the presence of an agent in the area where the object came from (suggesting that they inferred an agent must have thrown the object, as objects cannot generate their own motion; Saxe et al., 2005). The second goal of this thesis is to investigate the extent to which physical reasoning aids in our capacity to detect social information from physical objects (Chapter 3).

1.4 Inferring mental states from objects

Once we know that an object contains social information, we can make further social inferences about the agent that was involved. Humans possess a specialized cognitive system to process, understand, and predict each other's behavior in terms of their mental states, known as a *Theory of Mind* (Gopnik et al., 1997; Wellman, 2014). Equipped with this intuitive theory, people can infer the unobservable mental states that causally give rise to other people's observable behavior.

Despite the power of this cognitive capacity, it alone cannot explain how we make mental-state inferences from objects because the actions that normally serve as input for our Theory of Mind are not observable. One way to resolve this is to treat the object as a state in an agent's action sequence. For example, the stack of rocks from our opening vignette can be interpreted as the final action state in a sequence of actions where an agent stacked some number of rocks, one by one. Reasoning about how an agent's actions impact the physical world, and how objects interact with one another, requires another cognitive capacity: intuitive physics (Téglás et al., 2011; Spelke and Kinzler, 2007).

Though much is understood about these two cognitive capacities individually, very little is understood about how they work together to extract social information from physical objects. Here I briefly introduce a computational framework that instantiates a simplified version of these capacities. Computational models serve as a formal and exhaustive testing bed for scientific hypotheses, enabling us to test a full range of predictions across a full range of cognitively-inspired parameters. These results can then be directly compared with human data collected from behavioral experiments to evaluate the hypothesis.

Recent work has formalized Theory of Mind as Bayesian inference over a generative model of utility-maximizing action plans (Jara-Ettinger et al., 2016; Jara-Ettinger, 2019). That is, we first observe an action and attempt to sample mental states until we determine which set of mental states explains the action we observed. Specifically, the actions generated aim to maximize the agent's utility, adhering to the principle of rational action (Gergely and Csibra, 2003), and formalizing in precise probabilistic terms the essence of the previous qualitative approaches by Dennett (1989) and Gergely et al. (1995).

Mathematically, this work has expressed the problem of inferring an agent's beliefs b and desires d given some observed action a by:

$$p(b, d|a) \propto p(a|b, d)p(b, d)$$

where $p(a|b, d)$ captures how beliefs and desires lead to actions and $p(b, d)$ are an agent's prior assumptions about others' likely beliefs and desires. Given that actions cannot be observed in the examples I consider (Figure 1.2), I propose an extension to the traditional belief-desire psychology framework. Where agents normally reason about others' observable actions, in this extended framework they instead reason about observable traces in order to reconstruct the actions that led to them. This *event reconstruction* can then be used to infer others' goals, beliefs, and desires (see Figure 1.4). Here I focus on goal inference as a case study, though this model could be easily extended further to account for beliefs and desires (given another model that maps beliefs and desires to goals). Mathematically, this extended model can be expressed by:

$$p(g|t) \propto p(t|g)p(g)$$

where actions a are now replaced by observable traces t (e.g., an object). The term $p(t|g)$ can be further unpacked as follows:

$$p(t|g) = p(t|a)p(a|g)$$

The first term $p(t|a)$ represents our internal causal model of how actions leave behind traces in the environment (i.e., our intuitive physics). The second term $p(a|g)$ represents our internal causal model of how goals guide an agent's actions (i.e., our Theory of Mind). Given this extended model, we can now generate a probabilistic expectation of another agent's goals, all from interpreting an object as a trace of that agent's past actions. The third goal of this thesis is to investigate this model's efficacy in capturing human intuitions about what traces can reveal about the agents responsible for them (Chapter 4).

Given that an agent was involved in arranging a scene, and that we can reconstruct what happened, what kinds of mental states can we extract? The reconstruction does more than elucidate their likely goals. It can also tell us about the effort that went into arranging the scene. Consider again the examples in Figure 1.2. They all elicit a common mental state: communicative intent. Specifically, whoever arranged these objects had the intention to communicate with others that they should stay away, else why impose a cost? Critically, these objects do not truly restrict our actions—the imposed cost is negligible enough that we could ignore it, yet we do not because we understand what the other agent desires. This example highlights the phenomenon of *communicative objects*, which is one type of mental-state inference that we can make from objects.

We model this interaction using a social recursive reasoning framework similar to those used in pedagogical demonstrations (Ho et al., 2016; Shafto et al., 2014), pragmatics (Frank and Goodman, 2012; Goodman and Frank, 2016), and mental-state inferences (Ullman et al., 2009). Consider an initial scene s_0 of two exits in an office building. There is an agent, an enforcer, that does not want another agent, a decider, to walk through one of the doors. Indeed, this resembles almost all of the situations in Figure 1.2. The enforcer can place some number of objects in front of one of the doors to communicate their desires (to stay away), transforming scene s_0 into scene s. The enforcer's utility function can be described by:

$$U_E(s; a, s_0) = R_E(a)p_D(a|s) - C_E(s_0, s)$$

where $R_E(a)p_D(a|s)$ represents the enforcer's egocentric reward if the decider takes action a, weighted by the probability that they take that action, and $C_E(s_0, s)$ represent the enforcer's cost for generating scene s from scene s_0. Here we can see that the enforcer is incentivized to choose scene transformations that minimize the effort required to arrange them.

The decider simply chooses which door to walk through a, given the following

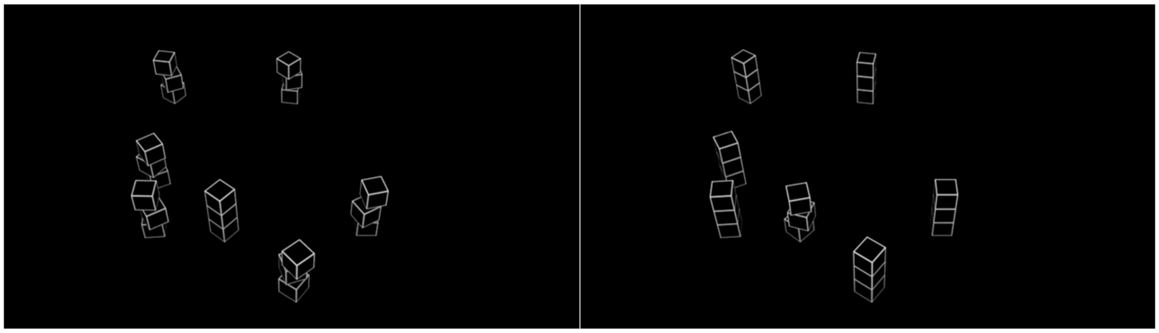

Figure 1.5: Example stimuli used in Experiment 1 (Chapter 2). On the left, there is a single perfectly-aligned tower in an array of slightly-misaligned towers. On the right, there is a single slightly-misaligned tower in an array of perfectly-aligned towers.

utility function:

$$U_D(a; s_0, s, \phi) = R_D(a) - C_D(a, s) + \phi < \ell(a|s_0, s) >$$

where $R_D(a)$ and $C_D(a, s)$ represents the decider's reward and cost for taking action a. $< \ell(a|s_0, s) >$ represents their "adopted utility" (Powell, 2022), weighted by a cooperation parameter ϕ that determines if they are cooperative, antagonistic, or apathetic towards the enforcer's desires. These utility functions reflect a recursive structure that is grounded at the implementation level. The central idea behind this proposal is that if people can infer other agents' mental states based on how they manipulated an object, then people can also strategically manipulate objects with the purpose of eliciting mental-state inferences in agents who encounter these objects. The fourth goal of this thesis is to test this model's efficacy in capturing human intuitions about the creation and interpretation of communicative objects (Chapter 5).

1.5 Thesis overview

1.5.1 Chapter 2

The existing literature on the perception of agency reveals how certain stimuli—those directly pertaining to agents—are prioritized in visual processing. Here I asked whether this might extend to physical objects. That is, does there exist low-level visual cues that reveal whether an object contains social information? I hypothesized that, moreso than perfect order, slight deviations from order is a potential cue. In a first experiment, I showed participants two types of block towers: one that consists of a perfectly-aligned stack of blocks, and another that consists of a stack of blocks that are slightly off from perfect alignment (see Figure 1.5). Using the visual search paradigm (inspired by the "stare-in-the-crowd" effect; Von Grünau and Anston, 1995), I found that participants were more accurate and faster at finding the block tower when it was misaligned (surrounded by an array of aligned towers) than when it was perfectly-aligned (surrounded by an array of misaligned towers). This result suggests that objects containing slight deviations from perfect order may be receiving attentional prioritization in visual processing.

1.5.2 Chapter 3

Another possible mechanism for detecting social information from physical objects is to consider whether an appropriate natural process could explain observed manipulations. I formalized this hypothesis as a computational model that infers the probability that an agent was involved in environmental manipulations by only considering the likelihood that the manipulations could occur by physics alone. In a first experiment, I showed participants synthetic images consisting of blocks arranged inside of a cardboard box with a hole cut out from the top, and they were tasked with determining if a particular arrangement was the result of an agent or a natural process (i.e., the blocks falling through the cutout). I found that this model generally captured participant intuitions about the plausibility of an agent's involvement. This result suggests that considering the physical plausibility of environmental manipulations is sufficient in detecting whether an object contains social information.

1.5.3 Chapter 4

Given the knowledge that an agent was involved in arranging a scene, we can now reason about them, despite never having encountered them. Here I proposed that the way people do this is through *event reconstruction*, where physical objects are seen as traces of an agent's previous actions, and can therefore be used to reconstruct them. In a series of experiments, I showed participants small gridworld representations of a room with multiple goals and entrances. In Experiment 1, I first tested whether our model matched human inferences in a task where participants had to infer an agent's entry point into the room and goal, all from a single pile of cookie crumbs that served as the trace. I found that the model strongly correlated with participant judgments, suggesting initial evidence for our event reconstruction account. In Experiment 2, I then explicitly tested people's ability to reconstruct the actions they believe different agents took based on indirect physical evidence of their presence. Here I also found that the paths that the model sampled tightly matched the paths that participants drew, lending further support to the idea that the inferences in Experiment 1 were supported by an ability to reconstruct events. In Experiment 3, I tested whether participants could infer whether one or two agents were involved, given two physical traces. I found that while this task was harder for participants, they were still able to infer the number of agents in each scene. Combined, these results suggest that event reconstruction supports how people infer social information from the physical traces that agents leave behind.

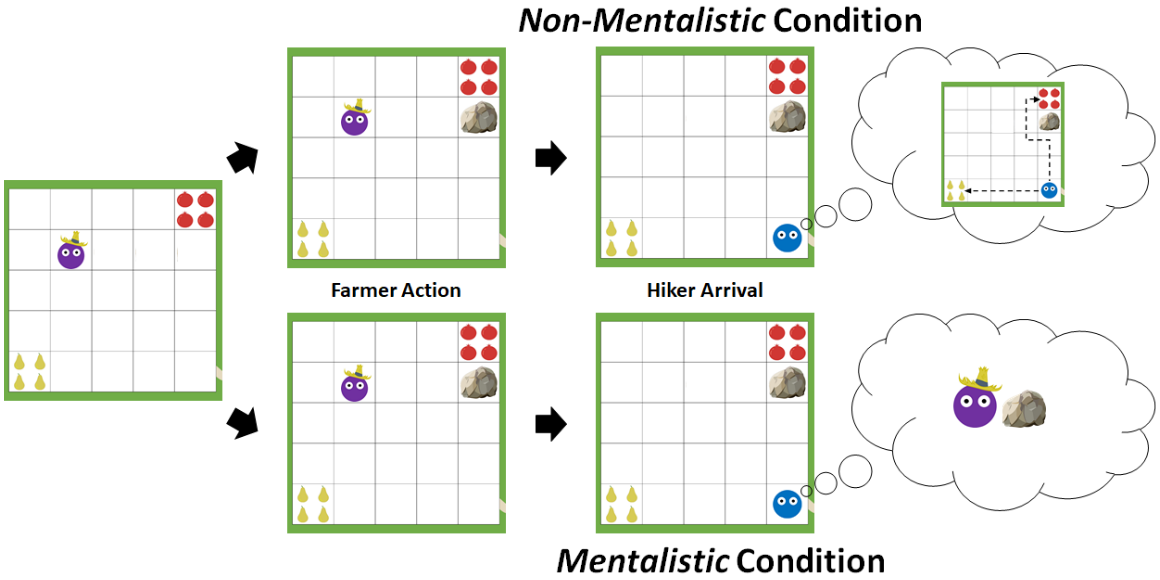

Figure 1.6: Visual schematic of our quantitative experiment cover story. Participants learned that a farmer (purple agent) wanted to protect their pomegranates and placed boulders to block the way before leaving. After leaving, a hiker would arrive and decide which fruit to take. In the *non-mentalistic* condition, the hikers treat the boulders as natural constraints, and they therefore decide what to do without thinking about the farmer. In the *mentalistic* condition, the hikers know that a farmer must have placed the boulders, and use this to infer what to do.

1.5.4 Chapter 5

Given that an agent was involved, and that we have an estimate of what they did, we can now infer their mental states. Here I focused on the phenomenon of *low-cost communicative blockers*, where agents use objects as communicative deterrents by imposing a small cost (e.g., Figure 1.2). I modified a Bayesian framework used to understand an agent's goals, desires, and beliefs from their actions to perform mental-state inference from physical objects. In Experiment 1, I presented a computational model that reveals that a combination of two intuitive causal models—Theory of Mind and a simplified version of intuitive physics—predicts participant responses in a graded inference task (see Figure 1.6). In Experiments 2-4, we present behavioral evidence against an account that suggests that people only rely on explicit pedagogy and convention to interpret these objects, rather than relying on any mental-state inference. I found that this model provided a strong quantitative fit to participant judgments. Furthermore, participant responses in our qualitative experiments aligned with our account, but not with the account based on explicit pedagogy and convention.

1.5.5 Chapter 6

Altogether, this work presents evidence towards a theoretical account of how we come to embed and extract social information from the physical world. In Chapter 6, I close with a discussion of my findings, and elaborate on some limitations and areas for future research.

Chapter 2

How do we detect the presence of agents?

2.1 Introduction

Our world is commonly carved into two domains: the physical and the social. From a pile of toy alphabet blocks, we can quickly perceive the color and shape of each individual block, the contrast between adjacent or nearby blocks, and even the stability of any stacked blocks (Cholewiak et al., 2013). Often what we are most interested in is not the blocks themselves, but the people building them. From minimal observation of others, we can perceive their dispositions without the slightest amount of thought, such as their trustworthiness and dominance (Todorov and Duchaine, 2008). However, even objects can contain rich social information—consider the alphabet blocks again, but arranged in a near-perfect stack. A once socially-devoid scene, it now reveals that an agent was involved. What about this scene makes this so obvious?

One possibility is that there exist low-level visual cues in the objects that agents interact with that reveal their involvement. This idea presumes that our visual system functions as a filter for these low-level visual cues—these *traces* of agency. By contrast, if these traces did not receive specialized visual processing, we would be overwhelmed by the sheer magnitude of objects in our visual environment. To resolve this, we prioritize stimuli through the selective operation known as attention, which acts as a filter to both what we perceive and what we think about (Chun, 2011; Treisman, 2006). The idea of specialized visual processing for certain types of stimuli is not a new idea. In fact, decades of work in vision science has shown that our visual system prioritizes the most social stimuli around: people. In particular, work on the perception of animacy has shown specialization over specific properties of agents, like faces (Kanwisher et al., 1997; Johnson et al., 1991; Simion et al., 2001) and body movements (Troje, 2013; Johansson, 1973, 1976; Peuskens et al., 2005). Beyond observing living, breathing bodies, even simple shapes appear as animate (Heider and Simmel, 1944) and intentional (Gao et al., 2010; van Buren et al., 2016) through the cue of self-propulsion (Premack, 1990; Hauser, 1998; Di Giorgio et al., 2017).

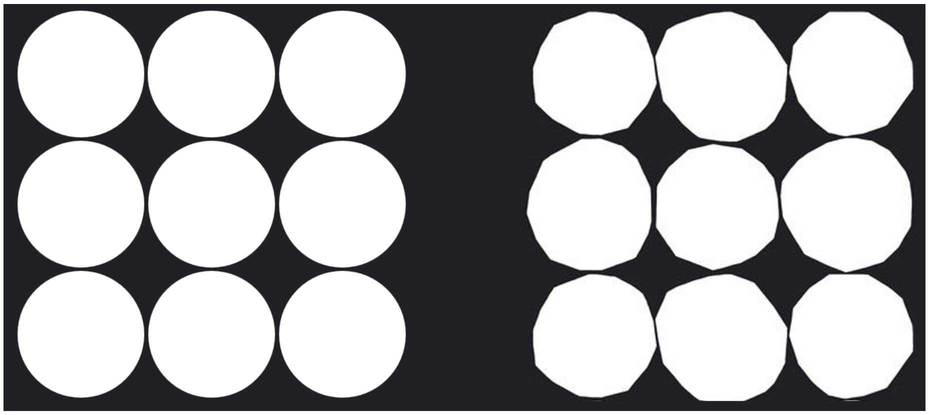

Figure 2.1: Example of a contrast between "perfect" order and order with slight errors. The grid of circles on the left appear so perfectly made, an agent would require a stencil to create. The grid of circles on the right, however, are just the kind of thing that an agent with a pair of scissors would produce.

This large body of work highlights the many ways in which the mind automatically and irresistibly extracts high-level properties from relatively simple visual stimuli, sometimes consisting of nothing more than a few shapes. If agents receive specialized visual processing, might some form of this specialization also extend to the traces of agency—such as the tower of alphabet blocks—and, if so, what are the cues?

Previous work has shown that from a very young age, children understand the link between agency and regularity (Newman et al., 2010; Keil and Newman, 2015a). Newman et al. (2010) showed that 12-month-old infants looked longer when an inanimate ball was responsible for creating an organized arrangement from a disorganized one, but did not exhibit preferential looking when an animate ball was responsible for the same action. In related work, 9-month-old infants associated a regular color sequence (e.g., yellow, yellow, red, yellow, yellow, red) as being the product of a human hand instead of a mechanical one (Ma and Xu, 2013). These findings reveal that even infants expect animate entities to be "creators of order" and inanimate entities to be "creators of disorder". This suggests that a potential cue for the traces of agency is whether or not an arrangement of objects contains regularity. In some cases, however, what is maximally regular sometimes seems artificial and non-agentive, and is instead ascribed to be the work of a machine or a supernatural entity (Aquinas, 1485; Dawkins et al., 1996). Suppose that Figure 2.1 is a black piece of construction paper with holes in it. The holes on the left seem unlikely to have been made by a person (without a specialized tool). The holes on the right—with many slight imperfections—seem like just the kind of outcome that we would expect from a person using a pair of scissors. While the link between regularity and agency is critical, a key part of this idea is that this link is *causal*. As we experience in our daily life, people are not infallible, so they leave behind traces of their actions, which manifest in the physical world as slight imperfections.

In this paper, we propose that these slight imperfections reveal if an arrangement of objects contains social information: specifically, whether or not an agent was involved. In Experiments 1a and 1b, we use the visual search paradigm (inspired by the "stare-in-the-crowd" effect; Von Grünau and Anston (1995)) to test participants' accuracy and speed when tasked with finding block towers that either slightly misaligned or perfectly aligned. Our proposal predicts that participants will have higher performance when searching for the slightly-misaligned block towers over the perfectly-aligned ones.

2.2 Experiment 1a

In Experiment 1a, we present an initial test of our hypothesis. In our main condition, participants were tasked with searching for one of two kinds of block towers: either slightly misaligned or perfectly aligned. If our proposal is correct, participants should exhibit better performance when looking for the slightly-misaligned block towers. One potential confound is that the collinearity of the adjacent blocks is driving the effect, rather than anything about block tower imperfections. To control for this, we ran a control condition where the blocks were no longer adjacent, disrupting their appearance as block towers, but still maintained their relative collinearity. Because spacing the blocks apart reveals additional contours (that were previously hidden when the blocks were stacked), we removed them to approximately match the visual complexity across both conditions. If participants are relying on collinearity, they should continue to perform similarly to participants in our main condition.

2.2.1 Participants

We recruited 200 English-speaking participants (with normal or corrected vision; $M = 32.57$ years, $SD = 12.73$ years) from Prolific.

2.2.2 Stimuli

Stimuli consisted of 1152 images (960 px × 540 px) of block towers made up of three blocks (made in Blender, version 2.79). Each block was black (#000000) with a white outline (#FFFFFF; stroke width = 7 px per block face). The background was also black (#000000) and was surrounded by a grey border (#808080). Half of these stimuli were for the main condition (e.g., Figure 2.2a), and were partitioned by several factors: the target block tower type (either slightly-misaligned, in an array of perfectly-aligned block towers, or vice versa), whether the target was absent or present (when absent, the target block tower was replaced by a block tower of the opposite type), whether the image contained seven or nine block towers. The other half of these stimuli were for the control condition (e.g., Figure 2.2c), had the same partitions, but were all vertically separated by an entirely black block (#000000) of equal size that matched the image background. Each misaligned block tower had a set of rotational offsets (sampled randomly from the set union of -35, -33, -31, ..., -15 and 15, 17, 19, ..., 35; measured in degrees) and a set of translational offsets (sampled randomly from the set -0.45, -0.4, -0.35, ..., 0.45; measured as a fraction of the block-width).

2.2.3 Procedure

Participants were randomly assigned to one of two between-subjects conditions: either the main condition or the control condition. Participants read a brief introduction to the task, containing examples of the two types of block towers they would be tasked to look for. Participants completed 12 experimental blocks, each containing 12 trials [two array sizes (7/9) × two target presence conditions (present/absent) × three repetitions], in a randomized order, for a total of 144 trials. Participants saw each image for 0.75 seconds (and preceded by a 0.50-second fixation consisting of the same black display except without any blocks). For each image, they were asked to press the ‘f’ key if the target block tower was present, and the ‘j’ key if it was absent.

2.2.4 Results

To measure participant performance, we computed their d' (a measure of sensitivity, that takes into account hits and false alarms; Green et al., 1966) and mean response time for each target block tower type (slightly-misaligned vs. perfectly-aligned). The first two trials of each experimental block were marked as practice, and data for these were not recorded. In our main condition, we found that participants were more accurate at finding the slightly-misaligned block towers ($M = 2.94$) than the perfectly-aligned ones ($M = 2.29$; $t(99) = 9.038$, $p < 0.001$ from a two-tailed t-test; Figure 2.2b, left). Participants were also faster at finding the slightly-misaligned block towers ($M = 0.83$) than the perfectly-aligned ones ($M = 0.90$; $t(99) = -7.531$, $p < 0.001$ from a two-tailed t-test; Figure 2.2b, right). In our control condition, participants were still more accurate at finding the slightly-misaligned towers ($M = 1.16$) than the perfect-aligned ones ($M = 0.90$; $t(99) = 4.083$, $p < 0.001$ from a two-tailed t-test; Figure 2.2d, left), and were also still faster ($M = 1.05$ vs. $M = 1.08$; $t(99) = -3.048$, $p = 0.003$ from a two-tailed t-test; Figure 2.2d, right). Despite there still being a performance advantage within the control condition, our primary question is whether this advantage is of the same magnitude as in our main condition (i.e., is the difference of differences significant). In this analysis, we found that the performance advantage in the main condition was significantly higher than that in the control condition, both in accuracy ($M = 0.65$ vs. $M = 0.27$, respectively; $t(196.176) = 3.947$, $p < 0.001$ from a two-tailed t-test) and response time ($M = -0.07$ vs. $M = -0.03$, respectively; $t(190.283) = -2.396$, $p = 0.018$ from a two-tailed t-test).

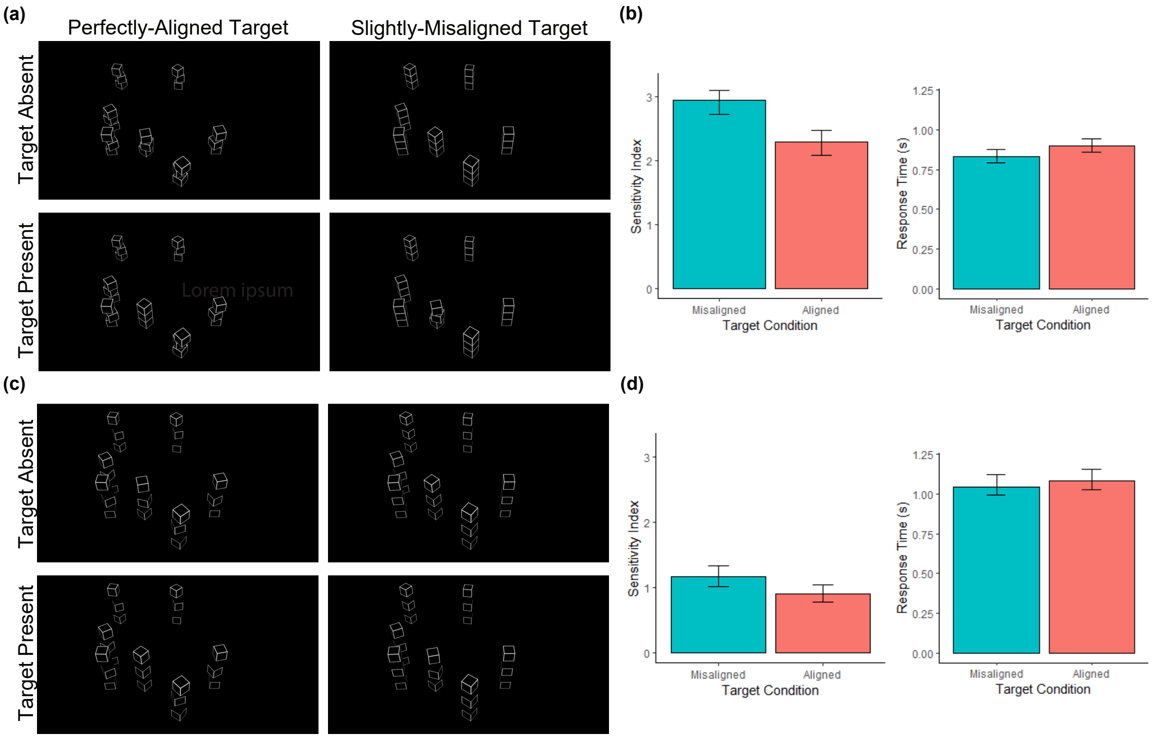

Figure 2.2: (a) Example stimuli configurations that participants saw in the main condition of Experiment 1a. (b) Mean sensitivity and mean response time for each target block tower type in the main condition of Experiment 1a. (c) Example stimuli configurations that participants saw in the control condition of Experiment 1a. (d) Mean sensitivity and mean response time for each target block tower type in the control condition of Experiment 1a. Error bars correspond to bootstrapped 95% confidence intervals.

2.3 Experiment 1b

In Experiment 1a, we compared our main condition against a control condition where the blocks were separated, in a manner that would approximately match the contours in the main condition. As a result, most of the blocks in these displays no longer resembled blocks. In Experiment 1b, we replicate our main condition from Experiment 1a and compare it against another control condition where the block towers continue to be separated, but no longer hide the additional contours that appear when moving the blocks.

2.3.1 Participants

We recruited 200 English-speaking participants (with normal or corrected vision; $M = 35.38$ years, $SD = 13.30$ years) from Prolific.

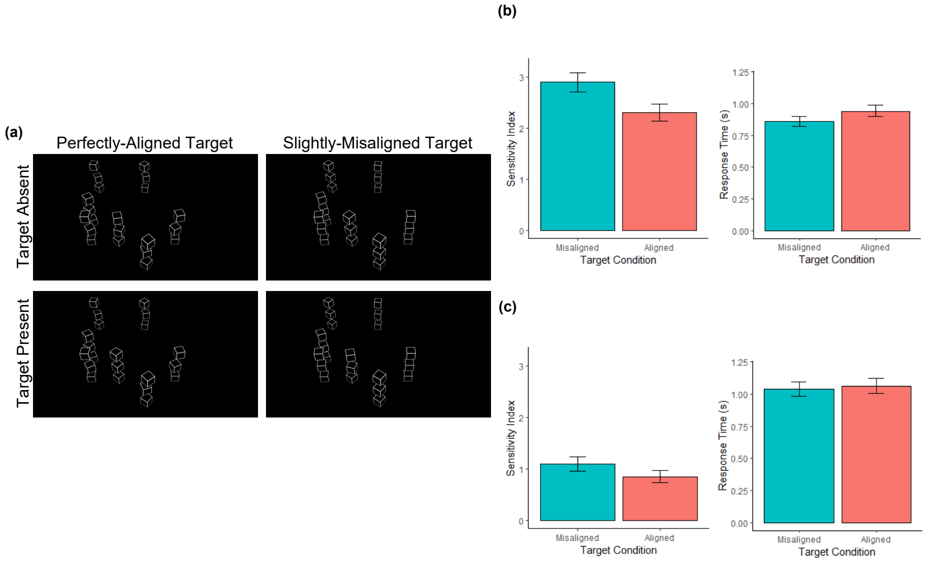

Figure 2.3: (a) Example stimuli configurations that participants saw in the control condition of Experiment 1b. (b) Mean sensitivity and mean response time for each target block tower type in the main condition replication of Experiment 1b. (c) Mean sensitivity and mean response time for each target block tower type in the control condition of Experiment 1b.

2.3.2 Stimuli

Stimuli consisted of 1152 images (960 px × 540 px) of block towers made up of three blocks (made in Blender, version 2.79). Each block was black (#000000) with a white outline (#FFFFFF; stroke width = 7 px per block face). The background was also black (#000000) and was surrounded by a grey border (#808080). Half of these stimuli were for the main condition, and were identical to those used in Experiment 1a. The other half of these stimuli were for the control condition (e.g., Figure 2.3a), had the same partitions, but were all vertically separated by an invisible block of equal size. Like in Experiment 1a, each misaligned block tower had a set of rotational offsets (sampled randomly from the set union of -35, -33, -31, ..., -15 and 15, 17, 19, ..., 35; measured in degrees) and a set of translational offsets (sampled randomly from the set -0.45, -0.4, -0.35, ..., 0.45; measured as a fraction of the block-width).

2.3.3 Procedure

The procedure was exactly the same as in Experiment 1a.

2.3.4 Results

As in Experiment 1a, we measured participant performance by computing their d' and mean response time for each target block tower type (slightly-misaligned vs. perfectly-aligned). The first two trials of each experimental block were marked as practice, and data for these were not recorded. In our main condition, we found that participants were more accurate at finding the slightly-misaligned block towers ($M = 2.91$) than the perfectly-aligned ones ($M = 2.31$; $t(99) = 8.391$, $p < 0.001$ from a two-tailed t-test; Figure 2.3b, left). Participants were also faster at finding the slightly-misaligned block towers ($M = 0.86$) than the perfectly-aligned ones ($M = 0.94$; $t(99) = -9.545$, $p < 0.001$ from a two-tailed t-test; Figure 2.3b, right). In our control condition, participants were still more accurate at finding the slightly-misaligned towers ($M = 1.09$) than the perfect-aligned ones ($M = 0.85$; $t(99) = 3.454$, $p < 0.001$ from a two-tailed t-test; Figure 2.3c, left), but were no longer faster ($M = 1.04$ vs. $M = 1.06$; $t(99) = -1.702$, $p = 0.092$ from a two-tailed t-test; Figure 2.3c, right). Again, our primary question is whether this advantage is of the same magnitude as in our main condition. In this analysis, we found that the performance advantage in the main condition was significantly higher than that in the control condition, both in accuracy ($M = 0.60$ vs. $M = 0.24$, respectively; $t(197.990) = 3.521$, $p < 0.001$ from a two-tailed t-test) and response time ($M = -0.08$ vs. $M = -0.02$, respectively; $t(167.135) = -3.658$, $p < 0.001$ from a two-tailed t-test). Combined with the results from our previous control condition, these findings suggest that the block towers are receiving some form of attentional prioritization due to their slight imperfections, rather than because of some low-level feature.

2.4 Discussion

Here we presented some initial evidence towards a cue-based account that proposes that we can detect the involvement of agents from the way they interact with objects in the environment. In Experiment 1a, we showed that participants exhibit stronger visual search performance for slightly-misaligned block towers over perfectly-aligned ones. In a control condition, we further showed that this could not be the result of collinearity between adjacent blocks, as separating the blocks—without introducing any new contours that would increase the visual complexity of the stimuli—led to participants' performance advantage to diminish. This control condition not only disrupted the perception of block towers (due to separating the blocks), but also led to most of the blocks no longer resembling blocks (due to removing some of the contours). In Experiment 1b, we replicated our main condition findings, and compared this replication with another control where we also separated the blocks, but this time did not hide the resulting contours, retaining the "objectness" of all blocks in the scene. Combined, these experiments suggest that our visual processing, specifically our attention, may be specialized for these kinds of traces of agency.

Our proposal revolves around "slight imperfections", but there are two possible ways to evaluate this, each importing its own assumption. The first way is to start with a goal state, and deviate slightly (e.g., by adding noise). For example, perceiving the imperfect holes in Figure 2.1 (right) as the product of an agent would involve either knowing or inferring the underlying goal state (i.e., Figure 2.1, left). The second way to evaluate an arrangement of objects for “slight imperfections” is to instead rely on an internal causal model of the world to simply reconstruct what is already in front of you. While both of these possibilities seem unlikely to occur online, given the sheer magnitude of computation that would be required, it is possible that they represent a computational-level description of our account. Establishing this connection is an open area of future work.

A related alternative hypothesis that we did not consider here is that these arrangements of objects may become salient through visual statistical learning (Turk-Browne et al., 2005). Consider seeing a stack of rocks in the forest. Our proposal would suggest that something about regularity makes this structure appear out-of-place, but it could also be that our visual system has “learned” that this arrangement is unlikely to occur in this context. The extent to which statistical learning plays a role remains an open question.

Vision is often associated with the perception of low-level features, such as colors and contrasts, but here we have expanded upon a rich body of work showing how vision can also extract social information, such as an agent’s involvement. We hope that this work will contribute to understanding the role that perception plays in helping us navigate the social world.

Chapter 3

Detecting social information through physical reasoning

This chapter is based on: Lopez-Brau, M., Kwon, J., McBean, B., Yildirim, I., & Jara-Ettinger, J. (2021). Detecting the involvement of agents through physical reasoning. *CogSci*.

Abstract

The physical world is rich with social information that people readily detect and extract, such as inferring that someone was present when we encounter a stack of rocks in the woods. How do people recognize that a physical scene contains social information? Research in developmental psychology has argued that this capacity is supported by a sensitivity to violations of randomness. Here we present a computational model of this idea and test its explanatory power in a quantitative manner. Our model infers agency by estimating the likelihood that a scene would arise naturally, as determined by human intuitive physics instantiated as a physics engine. Our results suggest that people's ability to detect agency in a physical scene is sensitive not only to the superficial visual properties, but also to the underlying physical generative process. Our results highlight how people use intuitive physics to decide when to engage in nuanced social reasoning.

3.1 Introduction

Human cognition is structured around our capacity to reason about the physical world—the spaces that we navigate and the objects we use and manipulate—and the social world—the agents that we encounter and the way we interact and cooperate with each other (Lake et al., 2017; Carey, 2009). A now large body of research suggest that reasoning about the physical and social world are supported by two initially distinct representational systems: Our naïve understanding of physics, which is grounded in representations of objects, their masses, and forces (Spelke, 1990; Baillargeon, 1987), and our naïve understanding of agency, which is grounded in representations of agents, their mental states, and intentional actions (Jara-Ettinger et al., 2016; Wellman, 2014; Woodward, 1998; Gergely and Csibra, 2003).

While these two systems support rich and flexible reasoning about objects and agents in isolation, many important problems in our everyday lives happen at their intersection. Understanding how agents act on the world often requires both representing what psychological states produced the agent's behavior (i.e., intuitive psychology), and how objects in the world will react to the forces that the agent's physical body applies (i.e., intuitive physics). Consistent with this, recent research suggests that the ability to integrate intuitive physics and intuitive psychology emerges early in childhood, supporting a range of social judgments, such as inferring the difficulty of different tasks (Gweon et al., 2017; Yildirim et al., 2019) and deciding when and how to help (Bennett-Pierre et al., 2018).

While these past studies show how people can combine social and physical reasoning when watching or interacting with agents, recent research suggests that this integration might also happen even in the absence of observable agents (Lopez-Brau and Jara-Ettinger, 2020; Schachner and Kim, 2018; Gosling et al., 2002; Lopez-Brau, 2021). For instance, people can infer what an agent was doing based on indirect evidence of their presence, such as a pile of breadcrumbs, and they can use the inferred actions to determine what the agent intended to do (Lopez-Brau, 2021; Lopez-Brau and Jara-Ettinger, 2020).

While these studies show that people can integrate intuitive psychology to interpret physical arrangements of objects, they do not shed light on how people determine when to engage in joint physical and social reasoning, particularly when no agents are present to suggest that social reasoning might be useful. One possible solution to this problem is that people first analyze physical environments through physical reasoning, and engage in social reasoning whenever they detect features of the environment that cannot be explained by physics alone. Consistent with this view, research suggests that violations of physics trigger an expectation of agency. For instance, after seeing an inert object fly into a scene, infants expect the presence of an agent in the area where the object came from (suggesting that they inferred an agent must have thrown the object, as objects cannot generate their own motion; Saxe et al., 2005). Similarly, infants infer the presence of an agent when a disordered set of objects becomes ordered but not the other way around (Newman et al., 2010).

The idea that people rely on a violation of natural physics to determine when social reasoning is necessary is consistent with the qualitative research reviewed above. Yet, to our knowledge, no work has sought to test this theory's explanatory power in a quantitative manner. In this paper we present a formal instantiation of this theory, implemented as a computational model that infers the involvement of an agent by considering the likelihood that the scene would occur by physics alone. To achieve this, we rely on advances in computational cognitive science suggesting that physical reasoning in humans is structured like a noisy *physics engine* (Battaglia et al., 2013; Gerstenberg et al., 2020) that supports mental physical simulations about objects and their forces. Here we propose that these same physical simulations can support the detection of agents. Critically, this account predicts that the detection of agents from physical scenes should not only depend on the arrangement of objects, but also on their physical properties. For instance, a set of blocks should give rise to different intuitions about the likelihood of an agent's involvement based on an observers' beliefs about the physical properties of the blocks (such as their weight) and the process they believed occurred before (such as the height at which the blocks were dropped). Thus, our experimental work focuses on this component: testing if people's intuitions about agency are affected by their beliefs about the causal process and the weight of the blocks, as well as by the likelihood that different ordered structures might appear naturally by physics.

3.2 Computational Framework

To make our focus concrete, consider the scenes shown in Figure 3.1. These displays represent a cardboard box with two cutouts. The cutout on the top of the box acts as a funnel and allows for blocks to fall through and land somewhere inside of the box. The cutout along the side of the box allows for an agent to manipulate the position of some or all of the blocks. Although we may never see the blocks falling through the funnel, we can intuitively reason that the arrangement of blocks in Figure 3.1a is unlikely to occur if the blocks were simply dropped into the funnel due to the height of the box, the fact that one of the blocks managed to stack on top of another, and the fact that this stack is quite far away from the funnel. Conversely, we would judge that the arrangement of blocks in Figure 3.1b is much more likely to occur naturally, since most of the blocks lie directly beneath the funnel and the block on the right may have simply bounced to that position. Our computational model aims to explain these intuitions.

We take as a starting point previous work showing that human naïve physics is instantiated as a physics engine that supports mental simulations about objects and their forces (Battaglia et al., 2013). In our model, however, rather than using a physics engine to predict the outcome of an event, we instead use it to infer the likelihood of a scene arising through some natural, physical process.

Formally, we model the environment (e.g., Figure 3.1a-b) inside of a physics engine

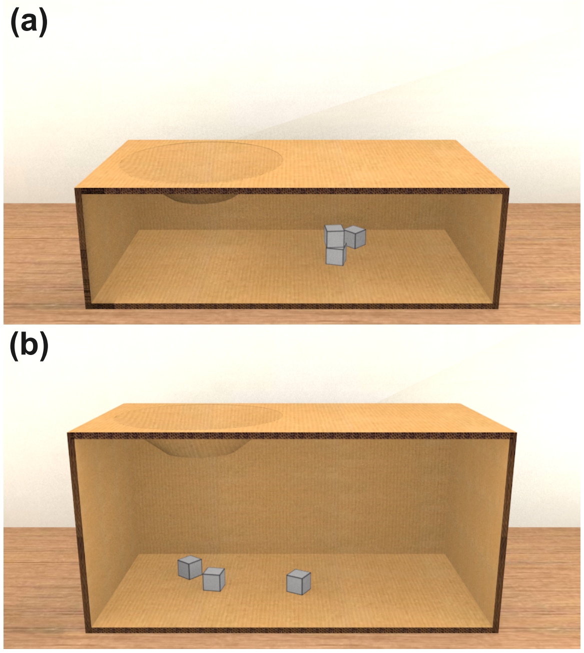

Figure 3.1: Examples of stimuli from our behavioral experiment. (a) Example of the short box stimuli. The blocks in this scene are arranged in such a way that it is implausible that they occurred naturally by falling through the funnel. (b) Example of the tall box stimuli. The blocks in this scene are arranged in such a way that it is plausible that they occurred naturally by falling through the funnel.

that simulates the process of dropping blocks through a funnel to build a distribution over expected physical outcomes. For simplicity, we first explain how we evaluate a scene with respect to a single simulation result from the physics engine, and then present how we achieve the full inference process.

Given a sample from the physics engine (i.e., the final outcome obtained from a physical simulation), we evaluate the observed scene by comparing its block arrangement against the block arrangement obtained from the sample. The simplest way to do this would be through a likelihood function that assigns a probability of 1 to the scene whenever the sample is an exact match, and 0 otherwise. Such a likelihood function, however, would be too strict, as it would fail to distinguish a sample that is close to the observed scene from a sample that is significantly different from the observed scene. To remove this concern, we assigned a non-zero probability to the scene when there were blocks in the scene that were positioned close to the blocks in the sample. To achieve this, we considered every possible permutation of blocks P that links the blocks in the scene to the blocks in the sample (i.e., every possible 1-1 correspondence relation), and then calculated the probability that each observed block b_j would appear as far as it did from the corresponding simulated block $\hat{b}_{p_j}$

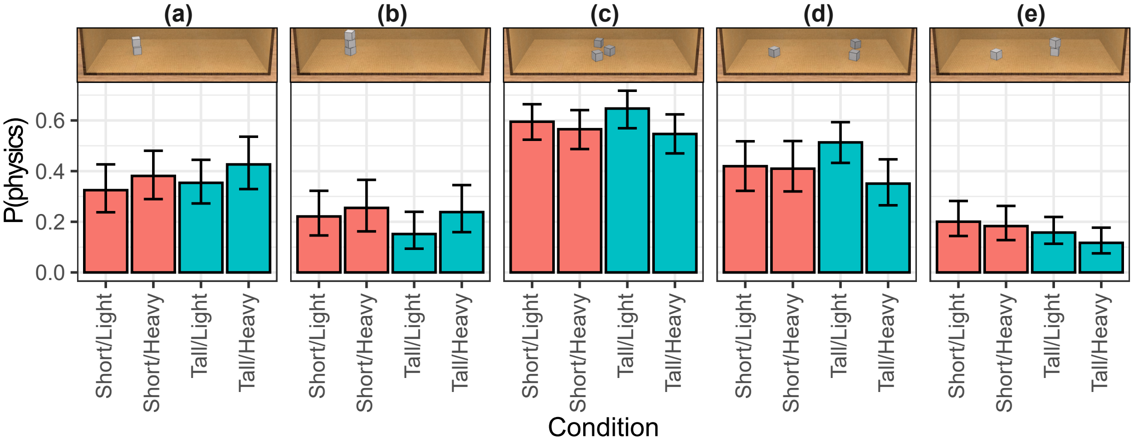

Figure 3.2: Subset of experiment results with corresponding stimuli above. (a-b) In these examples, participants judged them to be more likely with heavier blocks, since lighter blocks might be more prone to bouncing off one another. (c-d) In these examples, participants judged them to be more likely with a taller box, to give the blocks more leeway to bounce further into the box. (e) Since blocks stacking is already quite rare, seeing a stack of blocks away from the funnel leads participants to infer that an agent must have intervened. Following our color labeling from Figure 3.3, red bars correspond to the short box conditions and blue bars correspond to the tall box conditions. Errors bars are 95% bootstrapped CIs.

under each permutation $p \in P$, such that

$$p(b_j|\hat{b}_{p_j}) = \frac{p(X = d_{xy}(b_j, \hat{b}_{p_j}))p(Z = d_z(b_j, \hat{b}_{p_j}))}{p(X = 0)p(Z = 0)}$$

where $d_{xy}(b_j, \hat{b}_{p_j})$ represents the distance between the blocks in the xy-plane (computed by taking the difference between the radii of each observed block b_j and each simulated block $\hat{b}_{p_j}$ and, where (0, 0) is the point directly underneath the center of the funnel). This difference is then used to compute a likelihood by passing it as argument to a Gaussian probability density function (pdf) $P(X = d_{xy}(b_j, \hat{b}_{p_j}))$ and normalizing it by dividing by the max of this Gaussian pdf (i.e., $P(X = 0)$) in order to produce a valid probability. $d_z(b_j, \hat{b}_{p_j})$ is computed identically, except that the difference involves the z-coordinate of each block. Throughout, we set $X \sim \mathcal{N}(0, 7)$ to capture some noise tolerance on the xy-plane and $Z \sim \mathcal{N}(0, 0.01)$ to capture near zero tolerance on any perceived variation on the z-axis (thus preventing the possibility that a block stacked on top of another could be considered similar to a nearby block).

Given this method of calculating the probability that a sample $\hat{s}$ would produce the scene s under a mapping from simulated blocks to observed blocks p, we set the likelihood of the simulation generating the scene as

$$L(s|\hat{s}) = \max_{p \in P} \frac{1}{B} \sum_{j=1}^{B} p(b_j|\hat{b}_{p_j}).$$

This likelihood therefore captures the average likelihood of each simulated block matching an observed block, under the best association between simulated blocks and observed blocks.

Finally, we compute the final probability that a scene is generated by the physics engine through Monte Carlo sampling, such that

$$P(s|\text{physics}) \approx \frac{1}{N} \sum_{i=1}^{N} L(s|\hat{s}_i)$$

where $\hat{s}_i$ is a sample from the physics engine. Throughout, all model predictions were obtained by sampling 10,000 simulations.

3.3 Behavioral Experiment

We evaluated our model in an experiment where participants saw images like those shown in Figure 3.1 and they had to infer whether some or all of the blocks in the scene had been manipulated by an agent. Critically, our model predicts that these inferences should depend not only on the visible arrangement of the blocks, but also on the underlying physical process. We therefore manipulated both the weight of the blocks and the height of the box to test if this physical information would affect participant inferences, as our model predicts.

3.3.1 Participants

160 U.S. participants (as determined by their IP address) were recruited using Prolific (M = 34.58 years, SD = 12.20 years). 34 participants were excluded and replaced for failing our inclusion criteria.

3.3.2 Stimuli

Our stimuli consisted of 34 images of blocks arranged inside the bottom of a box, like those shown in Figure 3.1 (see `tinyurl.com/ylvkowh6` for full stimuli). Each image contained 1, 2, or 3 blocks, and blocks were always positioned around one of three locations: directly underneath the funnel, near the middle of the box, or on the distal end of the box opposite from the funnel. Our stimuli consisted of all permutations of these placements (including configurations where some or all of the blocks were stacked). We designed three trials with one block, nine trials with two blocks, and

22 trials with three blocks. Lastly, based on which of the two conditions participants were randomly assigned to, they would either see a short box (e.g., Figure 3.1a) or a tall box (e.g., Figure 3.1b). All stimuli were generated using the Blender 3D modeling software.

3.3.3 Procedure

Participants were randomly assigned to one of four conditions: light blocks inside of a short or tall box or heavy blocks inside of a short or tall box. This allowed us to test whether the exact same arrangement of blocks elicits different inferences as a function of participants' beliefs about the height of the box and the weight of the blocks. Participants read a brief introduction that explained the logic of the study, followed by a video that demonstrated how the blocks fall through the funnel and into the box. This video featured five blocks and a box of a different height than what participants would see in the trials, and served to give participants information about the weight of the blocks.

All participants completed the same 34 trials in a random order (varying only in the displayed box height, which depended on the condition). On each trial, participants were asked "How likely is it that a person moved some or all of the blocks?", using a continuous slider from "not likely" (i.e., 0) to "very likely" (i.e., 1). After the 17th trial, all participants were presented with a catch trial that was used as one of our exclusion criterion. This catch trial showed two blocks "glued" to the side of the box, in a way that is impossible to occur naturally, so people should judge that an agent must have manipulated the blocks.

3.3.4 Results

For each participant, their judgments were normalized by subtracting them by the participant's minimum judgment, and dividing this by the participant's maximum judgment minus the participant's minimum judgment. This same normalization procedure was applied to our model predictions for each condition. Finally, participant judgments were averaged together to produce a mean normalized participant judgment for each trial and condition. Figure 3.3 shows the results from Experiment 1. Overall, our model showed a correlation of $r = 0.86$ with participant judgments (95% CI: $0.82 - 0.89$). Within each condition, our model showed correlations of: $r = 0.87$ (95% CI: $0.79 - 0.91$) for a short box and light blocks; $r = 0.87$ (95% CI: $0.76 - 0.93$) for a short box and heavy blocks; $r = 0.89$ (95% CI: $0.84 - 0.93$) for a tall box and light blocks; and $r = 0.83$ (95% CI: $0.70 - 0.90$) a tall box and heavy blocks.

Figure 3.2 shows the results from five trials across the four conditions. As this figure shows, people's judgments show a subtle but highly systematic pattern across conditions that reveals their sensitivity to the physical properties of the blocks and the process that might have given rise to their arrangement (i.e., the height of the box and the weight of the blocks). Figure 3.2a shows a situation where a stack of two

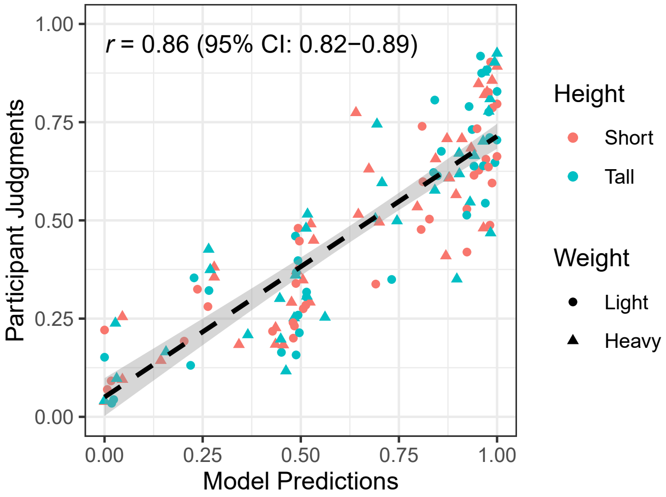

Figure 3.3: Results from our behavioral experiment. Each point corresponds to a judgment, with normalized model predictions on the x-axis and mean normalized participant judgments on the y-axis. Color indicates which condition the judgment corresponds to, and the fitted line shows the best linear fit with 95% confidence bands (in light grey).

blocks appeared directly underneath the funnel. Here, participants were more likely to believe this could have occurred naturally when the blocks were heavy relative to when they were light. Figure 3.2b shares an identical pattern, with higher probability when the blocks are thought to be heavier, but the additional stacked block leads to an overall decrease in probability across all conditions. Figure 3.2c-d show how judgments varied as a function of the box height. In Figure 3.2c, a set of three blocks that is farther away from the funnel is considered to be more likely when the blocks are light relative to when the blocks are heavy, with the highest inference when the blocks were light and the box was tall. In Figure 3.2d, the fact that two blocks are even farther away from the funnel lead to a main decrease in participant judgments, with the highest one corresponding to light blocks dropped from a tall box. Lastly, Figure 3.2e resembles Figure 3.2d except now the two blocks on the right are stacked. The overall decrease in the probability of this arrangement occurring naturally can likely be attributed to the implausibility of a block tower occurring so far from the funnel.

3.4 Discussion

Intuitive physics is often thought of as a capacity that helps us navigate the physical world. Our results show that this capacity can be used to do more than reason about the physical properties of the world: it can also be used to detect the involvement of agents in physical scenes. Here we explored the idea that the detection of agency from physical arrangements of objects is grounded in a sensitivity to violations of natural physics. We tested this possibility by implementing a computational model of human intuitive physics through a physics engine (Battaglia et al., 2013) and comparing it to human judgments in a range of parametrically-varying displays. Our experiment showed that people can infer the likelihood that an agent was involved a scene based on the possible physical process that could have given rise to the scene naturally.

Critically, under our account, people should be sensitive not only to the visual features of a scene, but also to their underlying beliefs about the physical process. Consistent with this, our model captured human judgments with quantitative precision in a task where we varied not only the arrangement of the blocks, but also participants' beliefs about the height at which the blocks were dropped and the weight of the blocks. Qualitative analyses of participant judgments showed a systematic structure that revealed sensitivity to the different physical features of each condition (height and weight). These effects, however, were subtle. One possibility is that this was due to a task artifact, as the conditions were run across participants and each participant therefore only ever saw one set of physical parameters (i.e., one box height and one block weight). In future work, we will replicate this study with a within-subjects design to better evaluate participant's sensitivity to physical properties.

Despite the high numerical fit, our experiment also revealed trials where participants and our model disagreed (see Figure 3.3). An initial qualitative analysis of these trials suggests that this was due to our model having a high noise tolerance when comparing blocks in a simulated sample to the observed blocks in a scene (by having a large variance in the Gaussian pdf for each block; Eq. 3.2). In current work we are increasing the number of physical simulations, enabling us to decrease the tolerance to mismatches between samples and observed scenes. This will allow us to obtain more precise inferences that will shed light on whether these model errors were due to a limitation in our inference procedure, or whether they reflect additional cognitive processes that humans use in our task but that are not captured in our model.

Here we focused on the inferences that people can make using only a model of physics. We believe such inferences are crucial for helping humans detect when to extract social information from physical scenes. For instance, hiking alone in a forest may involve little social reasoning until we encounter a scene that violates our expectations of what nature can do. Once we detect a violation of physics, people may engage in joint physical and social reasoning to explain what they see. It is possible, however, that social reasoning is much more pervasive than we recognize and that even physical arrangements of objects are always implicitly analyzed to see if they contain social information. If so, then a model that judges agency by simultaneously testing the physical plausibility of the scene (as our model does) and evaluating its consistency with what an agent would be likely and able to achieve, should outperform the current model that we presented. We are currently investigating this possibility.

Overall, our results suggest that human intuitive physics is central to human reasoning, not only because it enables us to reason about physical arrangements of objects, but also because its failure to explain what we see can help us determine when there is more than meets the eye. These results are a first step in characterizing how our intuitive theories enable us to seamlessly extract physical and social information from our surroundings.

Chapter 4

How do we reconstruct the past actions of agents?

This chapter is based on: Lopez-Brau, M., Kwon, J., & Jara-Ettinger, J. (2020). Social inferences from physical evidence via Bayesian event reconstruction. *Journal of Experimental Psychology: General*, *151*(9), 2029–2042. https://doi.org/10.1037/xge0001182

Abstract

Humans can make remarkable social inferences by watching each other's behavior. In many cases, however, people can also make social inferences about agents whose behavior they cannot see, based only on the physical evidence left behind. We hypothesized that this capacity is supported by a form of mental event reconstruction. Under this account, observers derive social inferences by reconstructing the agent's behavior, based on the physical evidence that revealed their presence. We present a computational model of this idea, embedded in a Bayesian framework for action understanding, and show that its predictions match human inferences with high quantitative accuracy. Specifically, Experiment 1 shows that people can infer where an agent came from and which goal they pursued in a room, all from a small pile of cookie crumbs. Experiment 2 shows that people can explicitly reconstruct the actions that the agent took, and these reconstructed trajectories can predict the entry point and goal inferences from Experiment 1. Finally, Experiment 3 shows that people can also infer whether one or two agents were in a room based on the position of two piles of cookie crumbs. Our results shed light on how people extract social information from the physical world.

4.1 Introduction

As social animals, humans possess a specialized cognitive system to process, understand, and predict each other's behavior, known as a *Theory of Mind* (Gopnik et al., 1997; Wellman, 2014). Theoretical and empirical work suggests that human Theory of Mind is instantiated as a mental model that specifies the causal relation between other people's unobservable mental states and their observable actions. That is, Theory of Mind captures how we expect other people's thoughts, preferences, and feelings to guide what they do. Equipped with this intuitive theory, people can infer the mental states that causally give rise to other people's observed behavior.

A rapidly growing body of work suggests that the causal model within Theory of Mind is structured around an assumption that agents act to maximize their utilities—the difference between the subjective costs they incur and the subjective rewards they obtain—capturing the idea that we intuitively expect others to act rationally and efficiently (see Jara-Ettinger, 2019, for review). Consistent with this view, computational models of mental-state inference via utility maximization reach human-level performance on simple social tasks (Baker et al., 2017; Jern et al., 2017; Jern and Kemp, 2015; Jern et al., 2011; Jara-Ettinger et al., 2020a), they capture richer forms of social behavior including pedagogy (Bridgers et al., 2020; Ho et al., 2019) and moral reasoning (Ullman et al., 2009), they explain social reasoning in early childhood and infancy (Gergely and Csibra, 2003; Jara-Ettinger et al., 2016; Liu et al., 2017; Lucas et al., 2014), and they have identifiable neural correlates (Collette et al., 2017).

Despite its success, this approach implicitly posits that mental-state inference requires access to someone's observable behavior, as it is these observed actions that enable us to evaluate the plausibility of different mental states. In many cases, however, people can make social inferences about agents whose behavior we did not get the opportunity to see. For example, imagine walking into an office building and finding a vacant receptionist desk with a chewed-up pencil, a half-filled crossword puzzle, and a cellphone. From this arrangement of objects, we can immediately infer that the receptionist might have been experiencing anxiety or restlessness (as the pencil was chewed-up), that they were likely procrastinating or had few tasks to complete at the moment (as they were working on a crossword), and that they expected to be gone only momentarily (as they chose to leave their valuable belongings unattended).

As the examples above show, human social inference is not limited to an ability to extract social information from observable actions—we can also make social inferences from physical scenes with no direct social or temporal information. How do we achieve this and how fine-grained are these inferences? Here we propose that social inferences about unobservable agents are supported by a basic form of *event reconstruction*, where, upon seeing indirect evidence of an agent's presence, we reconstruct what actions they likely took, enabling us to reason about the agent's behavior in a similar way to how we would if we had seen them act first-hand.

While it has long been known that the ability to infer social information from observed actions emerges early in infancy (Gergely and Csibra, 2003; Onishi and

Baillargeon, 2005; Woodward, 1998), recent studies suggest that social reasoning from physical events also emerges early in childhood. By preschool, children can estimate the difficulty associated with building different physical arrangements of objects (Gweon et al., 2017); they understand which kinds of actions leave physical traces in the environment and which kinds of actions do not (Jacobs et al., 2021); they can infer what someone knew based on physical evidence for how they searched an area (Pelz et al., 2020); and they can even detect the transmission of ideas by comparing artifacts created by different agents (Pesowski et al., 2020).

This past research suggests that the capacities needed to perform social inference via event reconstruction might be in place from childhood. However, to our knowledge, no work has formally explored the event reconstruction hypothesis that we propose here. Specifically, we hypothesized that people can causally reason about how goals lead to actions, and how actions leave traces in the environment. Combining these two causal models enables people to understand how goals lead to observable traces in the environments, connected by an inferred internal variable consisting of the actions that the agent took, which we call an event reconstruction. Here we present a computational model of this idea, testing social reasoning from agent-less physical scenes. Given indirect evidence that someone was present, our model infers what the agent was doing (i.e., reconstructs their actions) and why (i.e., infers their goals) through a generative model of how goals produce actions, and how actions leave observable evidence.

4.1.1 Connection to related proposals in social psychology

Consistent with our proposal, research in social psychology has found that people leave "behavioral residues" in their environments: physical cues that support rich inferences about their personality traits. For example, by looking at a picture of someone's messy desk, people can infer that the inhabitant is likely disorganized. From similar displays, people can also infer the inhabitant's degree of extraversion, conscientiousness, and even openness to new experiences (Webb et al., 1966; Gosling et al., 2002, 2008).

These inferences have been proposed to stem from a two-stage process, where people first use physical cues (such as a desk's cleanliness, the number of books in the room, or the cheerfulness of the décor) to infer someone's behavior, and then use this behavior to infer the underlying dispositions (Gosling et al., 2002; Brunswik, 1956). In this model, *cue utilization* captures how people transform these cues into social inferences, and *cue validity* captures whether these are accurate. Our hypothesis is consistent with this model, and it can be thought of as proposing that *cue utilization* consists of a form of Bayesian event reconstruction. From this standpoint, our work can be thought of as proposing a mechanism for how people associate different physical traces to the underlying behavior. Our work contributes to this literature by proposing a fully specified computational theory behind event reconstruction, grounded in the expectation that agents act rationally and efficiently in their environment, given their goals. Critically, however, previous models also account for inferences that people make based on stereotypes—a process that is outside of the scope of our work. We return to this point in the Discussion.

4.1.2 The current work

In Experiment 1, we first tested whether our model matched human inferences in a task where participants had to infer an agent's entry point into a room and their goal, all from a single pile of cookie crumbs that revealed their presence (see Figure 4.1). In Experiment 2, we then explicitly tested people's ability to reconstruct the actions they believe different agents took based on indirect physical evidence of their presence, lending further support to the idea that the inferences in Experiment 1 were supported by an ability to reconstruct events. Finally, if social reasoning from physical scenes is supported by event reconstruction, people should be able to also infer how many agents might have been present in a room, based on how many paths they need to reconstruct to explain the scene. We tested this prediction in Experiment 3. Combined, our results suggest that people have a nuanced capacity to infer social information from indirect evidence, and that these inferences are based on a basic capacity to "enhance" physical scenes by inferring agents' spatiotemporal behavior based on the indirect evidence that they leave behind. All studies were approved by the Yale University Institutional Review Board (protocol: "Online reasoning" #2000020357).

4.2 Computational Framework

Our model builds on a growing body of work showing that mental-state attribution is instantiated as Bayesian inference over a generative model of utility-maximizing action plans (Baker et al., 2009, 2017; Jara-Ettinger et al., 2020a; Jern et al., 2017; Jern and Kemp, 2015; Jern et al., 2011; Lucas et al., 2014). In our model, however, rather than evaluating unobservable goals against observable actions, we model how people might use physical evidence to reconstruct the actions that an agent took, and use these reconstructed actions to attribute goals.

To make our focus concrete, consider a situation like the ones shown in Figure 4.1a. Each of these displays represents a room with three possible goals (A in blue, B in orange, and C in green), two different doors (1 at the top in both rooms and 2 on the bottom and left, respectively), a set of walls (shown in dark gray), and a small pile of cookie crumbs that reveals that someone was previously in this room. Although we cannot see where this agent came from, what actions they took, or what goal they were pursuing, the cookie crumbs nonetheless contain information that we might be able to extract. In Figure 4.1a (left), the cookie crumbs intuitively reveal that the agent entered through door 1 and that they were likely pursuing goal A or C, but not goal B. In Figure 4.1a (right), the cookie crumbs intuitively reveal that the agent was pursuing goal C, but it is unclear whether they entered through door 1 or door 2. Our computational model aims to explain how we performed these inferences.

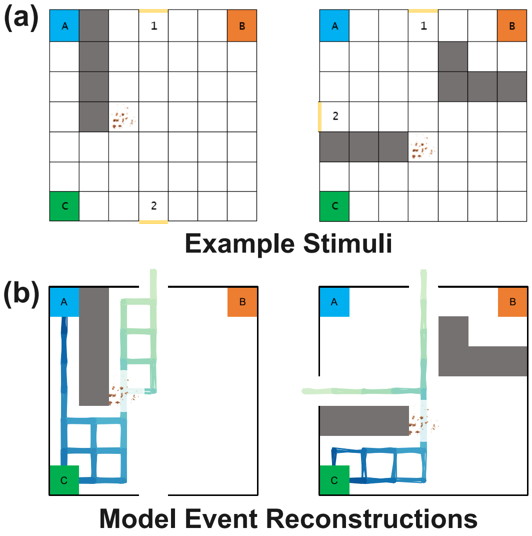

Figure 4.1: (a) Example stimuli from Experiment 1. Potential goals are positioned in the corners, labeled alphabetically, and color-coded. Doors are shown in yellow and coded numerically. Walls are shown in dark gray. Each trial included a pile of cookie crumbs positioned in a part of the room. (b) Visualizations of the underlying event reconstruction performed by our computational model for the examples above. Each line represents an inferred possible path, color-coded to indicate time, moving from light green to dark blue.

Formally, we model the environment as a gridworld, where the possible states of the world are given by the different positions in space that agents can occupy. At each time step, we assume that agents can move in any of the four cardinal directions and that these actions successfully move them in their intended direction (except when attempting to cross a wall, in which case the agent remains in the same position as they were before).

Given an observed static scene s (a gridworld with a set of goals, doors, walls, and a pile of cookie crumbs), the objective is to infer where the agent entered the room from (a door d) and which goal they pursued (a goal g), formally expressed as

$$p(d, g|s) \propto \ell(s|d, g)p(d, g),$$

where $\ell(s|d, g)$ is the likelihood of encountering scene s if an agent had indeed pursued goal g after entering through door d, and $p(d, g)$ is the prior over doors and goals.

According to our proposal, the ability to compute the likelihood function is mediated by a capacity to reconstruct the agent's actions. Under this view, if we can reconstruct the actions that the agent took, then judgments about the agent's entry point and goal are immediately revealed, as these are part of the reconstructed behavior (i.e., if we have access to the full reconstructed behavior, we can "see" where the agent entered from and where they were going). Formally, this idea can be implemented by expressing the likelihood function as

$$\ell(s|d,g) = \sum_{t\in\mathbb{T}} \underbrace{p(s|t)}_{\substack{\text{How do actions}\\\text{leave traces?}}} \times \overbrace{p(t|d,g)}^{\substack{\text{How do agents}\\\text{pursue goals?}}}.$$

Here $t = (\vec{s}, \vec{a})$ is a trajectory (from the set of all possible trajectories $\mathbb{T}$), which consists of an ordered sequence of pairs of states and actions that the agent took. $p(s|t)$ is the probability that an agent who took trajectory t would produce the observed scene s, and $p(t|g,d)$ is the probability that the agent would take trajectory t if they entered from door d with the intention to pursue goal g. This equation reveals the two components critical to our theory: an expectation of how agents navigate to complete their goals ($p(t|d,g)$), and an expectation of how agents' actions leave observable traces in the environment ($p(s|t)$).

To compute the expectations for how agents complete their goals, we used the standard framework previously developed in computational models of goal inference (Baker et al., 2009, 2017; Jara-Ettinger et al., 2020a) through Markov Decision Processes (MDPs)—a planning framework that makes it possible to compute the action plan or *policy* that maximizes an agent's utility function (Bellman, 1957). Classical MDPs are designed to produce a single trajectory that fulfills the agent's goal as efficiently as possible. In the cases that we consider, however, there are often multiple trajectories that can be equally efficient. As such, using a simple MDP can erroneously treat an efficient trajectory as unlikely if it is not an exact match to the solution that the MDP produced. To solve this problem, we built a probabilistic MDP that creates a probability distribution over all possible action plans, assigning higher probability to trajectories that are more efficient. Formally, we achieved this by softmaxing the MDP's value function when building the probabilistic policy. We used a low temperature parameter to identify all possible action plans that are equally (or approximately equally) efficient, enabling us to implement the expectation that agents navigate efficiently towards their goals. Using a probabilistic MDP, the probability that an agent would take trajectory t, starting from door d with the intention to fulfill goal g is given by

$$p(t|g,d) = \prod_{i=1}^{|t|} p(a_i|s_i,g),$$

where $p(a_i|s_i,g)$ is the probability of taking action a_i in state s_i, and the state sequence is given by trajectory t.

Finally, in our paradigm, we assume that the agent has a uniform probability of dropping the pile of cookie crumbs at any point in their path. The probability of observing scene s if the agent took trajectory t is therefore given by $p(s|t) = 1/|t|$ if the pile of cookie crumbs lies within the trajectory and 0 otherwise.

4.2.1 Implementation Details

To generate testable predictions, we set a number of parameters in our model prior to data collection. These choices are all reflected in our pre-registered model predictions (see `https://osf.io/q3ct5/`). We began by setting a uniform prior distribution over doors and goals, such that agents were equally likely to enter through any of the doors and equally likely to pursue any of the goals. Next, to model the forces that shape agents' actions, we assumed that agents incur a constant cost of 1 for any action that they take, and that goals produced numerical rewards over the range $0-100$. Finally, to make our MDP probabilistic, we applied a temperature parameter $\tau = 0.15$ to the value function. This parameter was set *a priori* to ensure that the model would give equal probability to all paths that were equally efficient, while only placing a negligible probability on erroneous and inefficient trajectories.

Model inferences were obtained via Monte Carlo methods, sampling 1000 combinations of doors and goals and 1000 trajectories conditioned on the selected door and goal. Figure 4.1b visualizes our model's inferred trajectories for the examples shown in Figure 4.1a, with each line corresponding to a sample from the posterior distribution, color-coded to indicate time, moving from light green to dark blue. These visualizations show how our model reconstructs the agent's probable spatiotemporal behavior, which in turn reveal the agent's entry point and goal, matching the intuitive inferences associated with these examples in the introduction.

4.3 Experiment 1a

In Experiment 1a, we tested our model in a task where people had to infer which goal an agent was pursuing and where they came from, all from a single piece of indirect evidence about their presence. If people's ability to infer goals from physical evidence is mediated by event reconstruction, then their judgments should show a quantitative fit to our model predictions, including fine-grained patterns of uncertainty. This study was pre-registered; all study materials can be found at `https://osf.io/q3ct5/`.

4.3.1 Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 37.02$ years, $SD = 11.20$ years).

4.3.2 Stimuli

Stimuli consisted of 23 gridworld images, like those in Figure 4.1a. Each gridworld was 7-by-7 squares in size and represented a room that contains three goal squares (A in blue, B in orange, and C in green), up to three doors (labeled 1, 2, and 3), and a pile of cookie crumbs. The goals were always in the same corners, but the position of the doors and the pile of cookie crumbs varied between trials. In addition to these three features, a subset of trials included walls (shown by the dark gray squares in Figure 4.1a) that agents could not walk through.

Our stimuli set was designed to capture different types of inferences while also controlling for features that simple heuristics could exploit (e.g., ensuring that the target goal was not always the one closest to the cookie crumbs, and that it could not be determined by projecting a straight line that intersected the entrance and the location of the cookie crumbs). We began by considering four different possible inference patterns: assigning probability close to 1 to a hypothesis (HIGH CERTAINTY trials), assigning probability close to 0 to a hypothesis, while also not having full certainty over two remaining hypotheses (HIGH NEGATIVE CERTAINTY trials), assigning a higher probability to one of the hypotheses (PARTIAL CERTAINTY trials), and assigning an approximately uniform distribution to the hypothesis space (UNCERTAIN trials).

We first designed seven single-door trials that captured each of these inference patterns in goal inference (two HIGH CERTAINTY, HIGH NEGATIVE CERTAINTY, and PARTIAL CERTAINTY trials, and one UNCERTAIN trial; schematic versions shown in Figure 4.3a). We then designed 16 additional trials with multiple doors by combining every possible inference pattern for the goal the agent was pursuing and the entrance that they took (schematic versions shown in Figure 4.3b).

4.3.3 Procedure

Participants read a brief tutorial that explained the logic of the task. After learning how to interpret the images, participants were told that agents were equally likely to enter the room from any of the doors with the intention of going directly to one of the three goals (to remove the possibility that agents pursue multiple goals, or wander aimlessly before selecting one). After the introduction, participants completed a questionnaire that ensured they read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 23 trials in a random order. On each trial, participants answered a multiple-choice attention-check question (“Which corner is farthest from Door 1 (there may be more than one)?”) and were asked to infer the agent’s goal (“Which corner is the person going for?”) using three continuous sliders, one for each goal (each ranging from 0, labeled as “definitely no,” to 1, labeled as “definitely”). Trials with at least two doors included a third question that asked participants to infer the agent’s entry point (“Which door did they come from?”) using one slider per door (each also ranging from 0, labeled as “definitely no,” to 1, labeled as “definitely”). Participants were allowed to submit their responses for each trial only when they correctly answered the attention-check question. Otherwise, participants were prompted to “please pay attention and try again.”

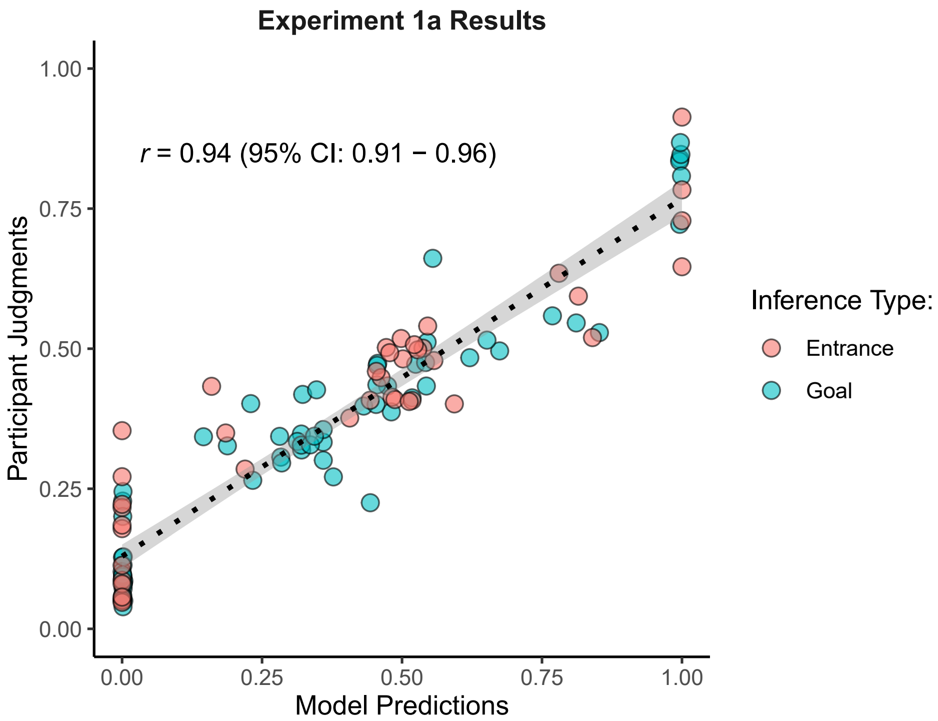

Figure 4.2: Results from Experiment 1a. Each point corresponds to a judgment, with model predictions on the x-axis and mean participant judgments on the y-axis. Color indicates inference type and the dotted line shows the best linear fit with 95% confidence bands (in light gray).

4.3.4 Results

Each participant's judgments were first normalized within-trial (such that every distribution over goals or doors added up to 1) and then averaged across participants. Figure 4.2 shows the results from Experiment 1a. Overall, our model showed a correlation of $r = 0.94$ (95% CI: $0.91 - 0.96$) with participant judgments, and the strength of the model fit was similar when looking only at goal inferences ($r = 0.95$; 95% CI: $0.92 - 0.97$) or door inferences ($r = 0.92$; 95% CI: $0.86 - 0.95$).

Figure 4.3 shows our model's results as a function of trial. In each subplot, the image at the top shows an abstract schematic of the trial, with the pile of cookie crumbs marked as a brown square. This figure reveals how our model not only predicted participant judgments in situations where the agent's entry point and goal were clear, it also matched participant judgments in its expression of uncertainty. Critically, our model produced nuanced patterns of uncertainty across trials, which reflect how well it was able to reconstruct the event, becoming less confident as a function of how much conflict there is in entry points and goals across different hypothetical reconstructions. The fact that this event-based uncertainty matched participant judgments with quantitative accuracy suggests that participants may have also been performing these inferences via some form of event reconstruction.

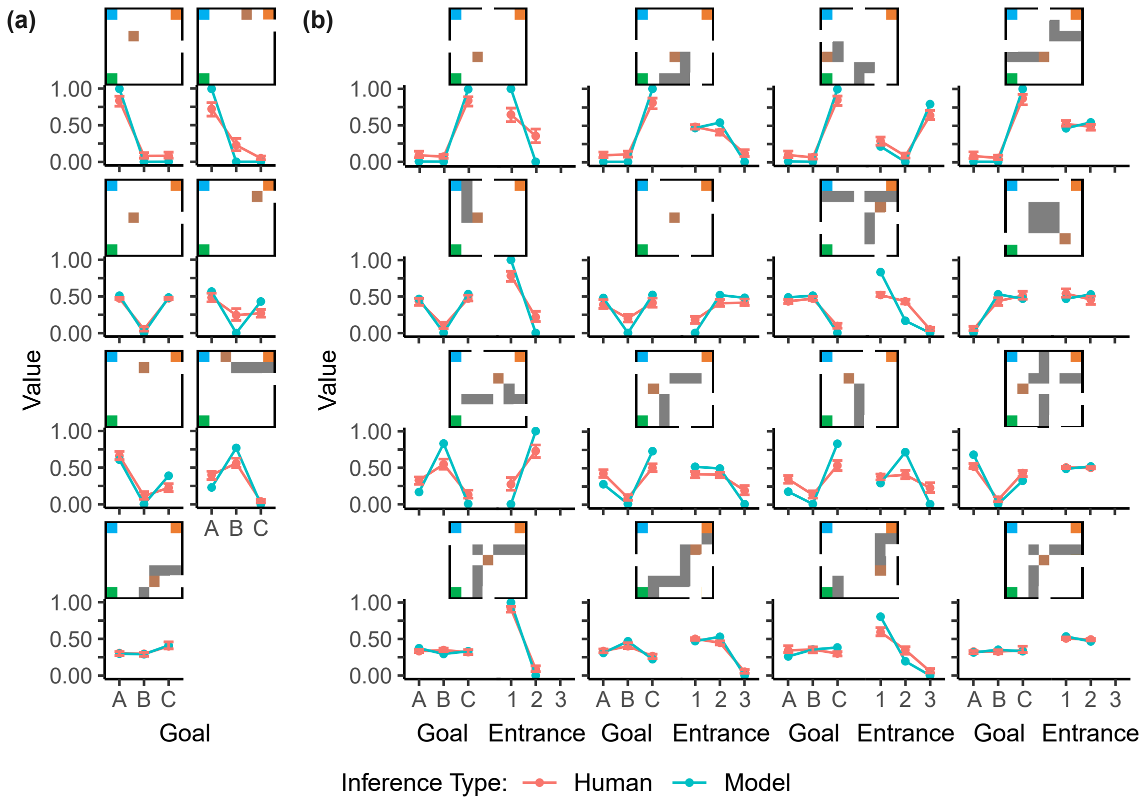

Figure 4.3: Detailed results from Experiment 1a. From top to bottom, each row of subplots corresponds to the HIGH CERTAINTY, HIGH NEGATIVE CERTAINTY, PARTIAL CERTAINTY, and UNCERTAIN trials for goal inferences, respectively. (a) Results for trials that only had one door. (b) Results for trials that had more than one door. From left to right, each column of subplots corresponds to the HIGH CERTAINTY, HIGH NEGATIVE CERTAINTY, PARTIAL CERTAINTY, and UNCERTAIN trials for door inferences, respectively. The goals A, B, and C are indicated by the blue, orange, and green squares, respectively. The doors are sequentially numbered in a clockwise fashion, with door 1 starting from the top (or from the right if there is no top door). Walls are marked as dark gray squares and the pile of cookie crumbs are indicated by the brown squares. Red lines represent mean participant judgments and blue lines represent our model's predictions. Error bars on participant judgments represent 95% bootstrapped confidence intervals.

One possibility is that the underlying goals or entry points of the agent correlate with superficial features of the stimuli, such as the proximity of the cookie crumbs to different doors or goals. If this is the case, then participants may have been able to infer agents' entry points and goals without performing any form of event reconstruction. We tested this possibility through a multinomial logistic regression trained to predict participant goal inferences as a function of the distance between the pile of cookie crumbs and each goal, the average distance between the pile of cookie crumbs and each door, the number of doors, and all of their interactions. To train this model, we transformed participant judgments into a one-hot vector, marking 1 for the goal with the highest probability and 0 for the rest, and implemented LASSO regularization (Tibshirani, 1996) to avoid overfitting. We generated the alternative model's predictions in a leave-one-out fashion—that is, the predictions for each trial consisted of the output of a regression trained on all remaining trials.

Even though this alternative model was trained on the qualitative structure of participant judgments, it nonetheless only produced a correlation of $r = 0.49$ (95% CI: $0.30 - 0.63$) with participant judgments, which was substantially lower than the one produced by our model ($\Delta r = 0.46$; 95% CI: $0.33 - 0.65$). These results show that, while superficial features can capture the broad structure of participant judgments, they fail to do so at our model's level of granularity, further suggesting that people's inferences were centered on a form of Bayesian event reconstruction.

4.4 Experiment 1b

Experiment 1a showed initial evidence for our model in a situation where people had no prior information about the agent. In many cases, however, people have prior knowledge about others, and this information affects their inferences. In Experiment 1b, we therefore tested if our model continued to capture participant inferences in a context where people were given prior information about the agent's behavior. This study was pre-registered; all study materials can be found at `https://osf.io/q3ct5/`.

4.4.1 Participants

160 English-speaking participants were recruited using Prolific ($M = 33.49$ years, $SD = 11.36$ years).

4.4.2 Stimuli

Stimuli consisted of 16 gridworld images, evenly divided across a *door prior* and a *goal prior* condition. Each gridworld was similar to those in Experiment 1a, with the difference that each trial now included prior information about an agent's behavior. In the *door prior* condition, each gridworld contained nine red 'X' markers, distributed across the doors. These markers represented the number of times the agent previously entered through each door. In the *goal prior* condition, each gridworld contained nine red 'X' markers, distributed across the three goals. These markers represented the number of times the agent previously pursued each goal.

To construct the stimuli for the *goal prior* condition, we first selected four gridworlds from Experiment 1a's PARTIAL CERTAINTY condition, and four gridworlds from Experiment 1a's UNCERTAIN condition (with respect to goal inferences). For each selected gridworld, we considered four possible prior distributions over the goals: $\{(3, 3, 3), (6, 2, 1), (1, 6, 2), (2, 1, 6)\}$. Because this condition consisted of eight gridworlds, each possible prior distribution was randomly assigned to one gridworld from the PARTIAL CERTAINTY set and to one gridworld from the UNCERTAIN set. This assignment was randomized across participants to ensure an equal amount of data for every possible combination of gridworld and prior distribution (resulting in a total of $8 \times 4 = 32$ possible combinations).

The stimuli for the *door prior* condition were designed in a parallel way. We first selected four gridworlds from Experiment 1a's PARTIAL CERTAINTY condition, and four gridworlds from Experiment 1a's UNCERTAIN condition (this time with respect to door inferences). Because all gridworlds from the PARTIAL CERTAINTY condition had three doors, we used the same set of priors and assignment procedure used in our *goal prior* condition described above. By contrast, all gridworlds from the UNCERTAIN condition had two doors. The priors for these trials were therefore sampled from the set $\{(5, 4), (5, 4), (7, 2), (2, 7)\}$.[1]

4.4.3 Procedure

The procedure was nearly identical to Experiment 1a, except that participants were also taught how to read the prior information. Participants were told that, in each gridworld, they would see the agent's entry point or goal (depending on condition) for the agent's nine previous visits, and their task was to infer the agent's entry point and goal for the tenth event. After the introduction, participants completed a questionnaire that ensured they read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 16 trials in two experimental blocks, one for the *door prior* condition and another for the *goal prior* condition. Experimental block order and within-block trial order were randomized across participants. The prior information on each trial was determined by one of four distributions (see Stimuli). On each trial, participants answered a multiple-choice attention-check question ("Which corner is the farthest walk from Door 1? If there is more than one correct answer,

[1] The pre-registered duplication of (5, 4) in the prior set was accidental, as it was meant to be (4, 5). This affected only 4 of the 64 possible gridworld-by-prior tests, and our experiment continues to have the necessary variability to compare participants to our model.

just choose one of them.") and were asked to infer the agent's goal ("Which corner is the person going for?") using three continuous sliders, one for each goal (each ranging from 0, labeled as "definitely no," to 1, labeled as "definitely"), and asked to infer the agent's entry point ("Which door did they come from?") using one slider per door (each also ranging from 0, labeled as "definitely no," to 1, labeled as "definitely"). Participants were allowed to submit their responses for each trial only when they correctly answered the attention-check question. Otherwise, participants were prompted to "please pay attention and try again."

4.4.4 Model Predictions

Model predictions were obtained in the same way as Experiment 1a, with the difference that the prior distribution over goals and doors was based on agents' prior behaviors. To achieve this, we began with a uniform distribution over goals and doors for every gridworld, and updated each distribution through Bayes' rule based on the prior behavior (i.e., the nine observations) shown in the gridworld, using the generative process specified in our model (i.e., by assuming that agents probabilistically choose the goal with the highest utility, subject to a softmax process with temperature $\tau = 0.1$). The resulting distributions were then set as the prior distributions in the study.

4.4.5 Results

Data was analyzed in the same way as Experiment 1a. Each participant's judgments were first normalized within-trial (such that every distribution over goals or doors added up to 1) and then averaged across participants for each condition. Figure 4.4 shows the results from Experiment 1b. Overall, our model showed a correlation of $r = 0.91$ (95% CI: $0.89 - 0.92$) with participant judgments, and the strength of the model fit was similar for the *goal prior* condition ($r = 0.91$; 95% CI: $0.89 - 0.93$) and the *door prior* condition ($r = 0.90$; 95% CI: $0.86 - 0.92$). Critically, these inferences once again revealed that participants produce graded patterns of confidence across trials, as predicted by our model. Together, these results show that people, like our model, can integrate prior information about how an agent behaved to reconstruct their actions given indirect physical evidence.

4.5 Experiment 2

In Experiment 1, we found that people can infer where an agent was going and where they came from, all from a single piece of indirect evidence about their presence. Participant judgments were quantitatively predicted by a model centered on an ability to reconstruct what happened. If our account is correct, then people should also be able to explicitly reconstruct the actions that an agent took in a way similar to our

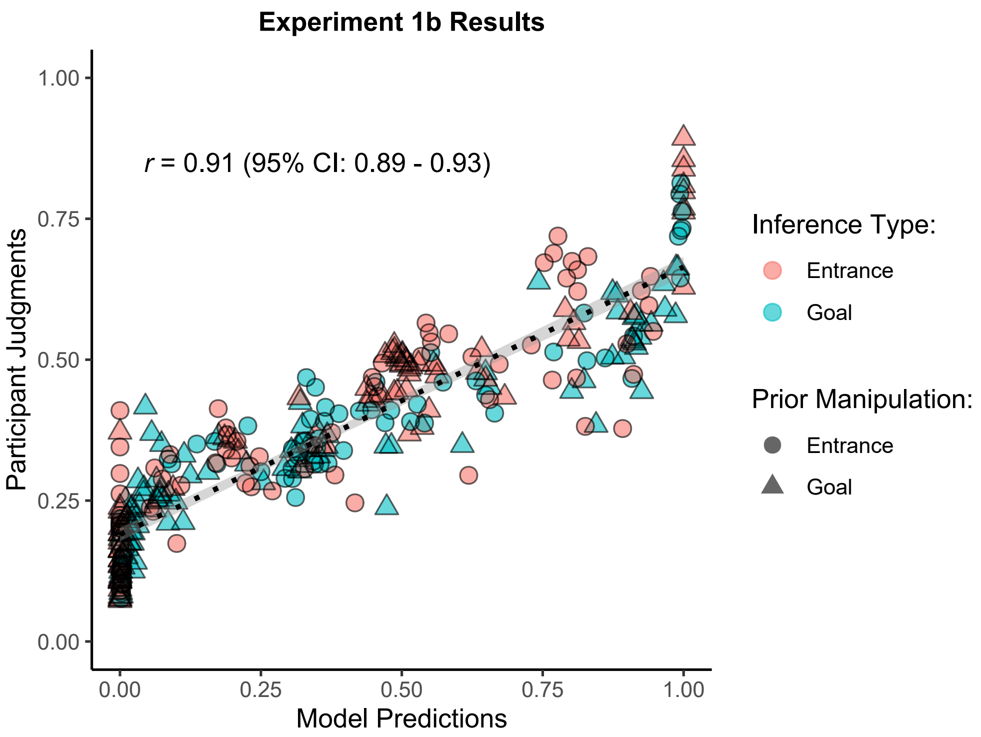

Figure 4.4: Results from Experiment 1b. Each point corresponds to a judgment, with model predictions on the x-axis and mean participant judgments on the y-axis. Color indicates inference type, shape indicates condition, and the dotted line shows the best linear fit with 95% confidence bands (in light gray).

model. We test this prediction in Experiment 2. This study was pre-registered; all study materials can be found at `https://osf.io/q3ct5/`.

4.5.1 Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 38.25$ years, $SD = 11.02$ years).

4.5.2 Stimuli

The stimuli were the same as those from Experiment 1a (see Figure 4.1a for examples and Figure 4.3 for schematic versions).

4.5.3 Procedure

Participants read a brief tutorial that explained the logic of the task. Participants were then taught how to draw their paths. After the introduction, participants completed a questionnaire that ensured they read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 23 trials in a random order. On each trial, participants were asked to infer the path they thought the agent took, given the pile of cookie crumbs. Participants generated their paths by sequentially clicking on the squares they believed the agent walked through. Participants were only allowed to proceed when they had successfully generated a valid path, which consisted of paths that started at a door, ended at a goal, and passed through the pile of cookie crumbs. Participants were allowed to reset the drawn path as many times as they wished.

4.5.4 Model Predictions

To evaluate the participant-generated path reconstructions, we used our framework to calculate

$$p(t|s) \propto p(s|t)p(t),$$

where $p(s|t)$ is the likelihood of a trajectory t generating scene s and $p(t)$ is the prior over possible trajectories. Here, $p(s|t) = 1/|t|$ (like in Equation 4.2) and $p(t)$ is obtained by marginalizing over agents' potential goals and entry points, as follows:

$$p(t) = \sum_{d,g} p(t|d,g)p(d,g).$$

4.5.5 Results

Our computational framework enables us to calculate the probability assigned to each path generated by participants. However, directly interpreting these probabilities is difficult, as they are sensitive to the length of the path and to the number of competing paths that fulfill a goal efficiently. To make our results easier to interpret, we compared our model's evaluations of the participant-generated path reconstructions with that of a baseline model. This baseline model used a uniform transition function over all actions, excluding the one that would generate a transition to the previous state (to prevent infinite back-and-forth loops). For every participant, we computed the Bayes factor for each of their reconstructed paths by dividing the probability of each path, as predicted by our model (i.e., $p(t|s)$), by the probability predicted by the baseline model. A Bayes factor greater than one would indicate that our model explains a participant's reconstructed path better than the baseline model; a Bayes factor less than one would indicate that the baseline model explains a participant's reconstructed path better than our model.

Our model outperformed the baseline model on all trials. The average Bayes factor in our experiment was 16935.33 (lowest factor = 7933.79; highest factor = 84383.12), meaning that our model was, on average, much more likely to produce the participant-

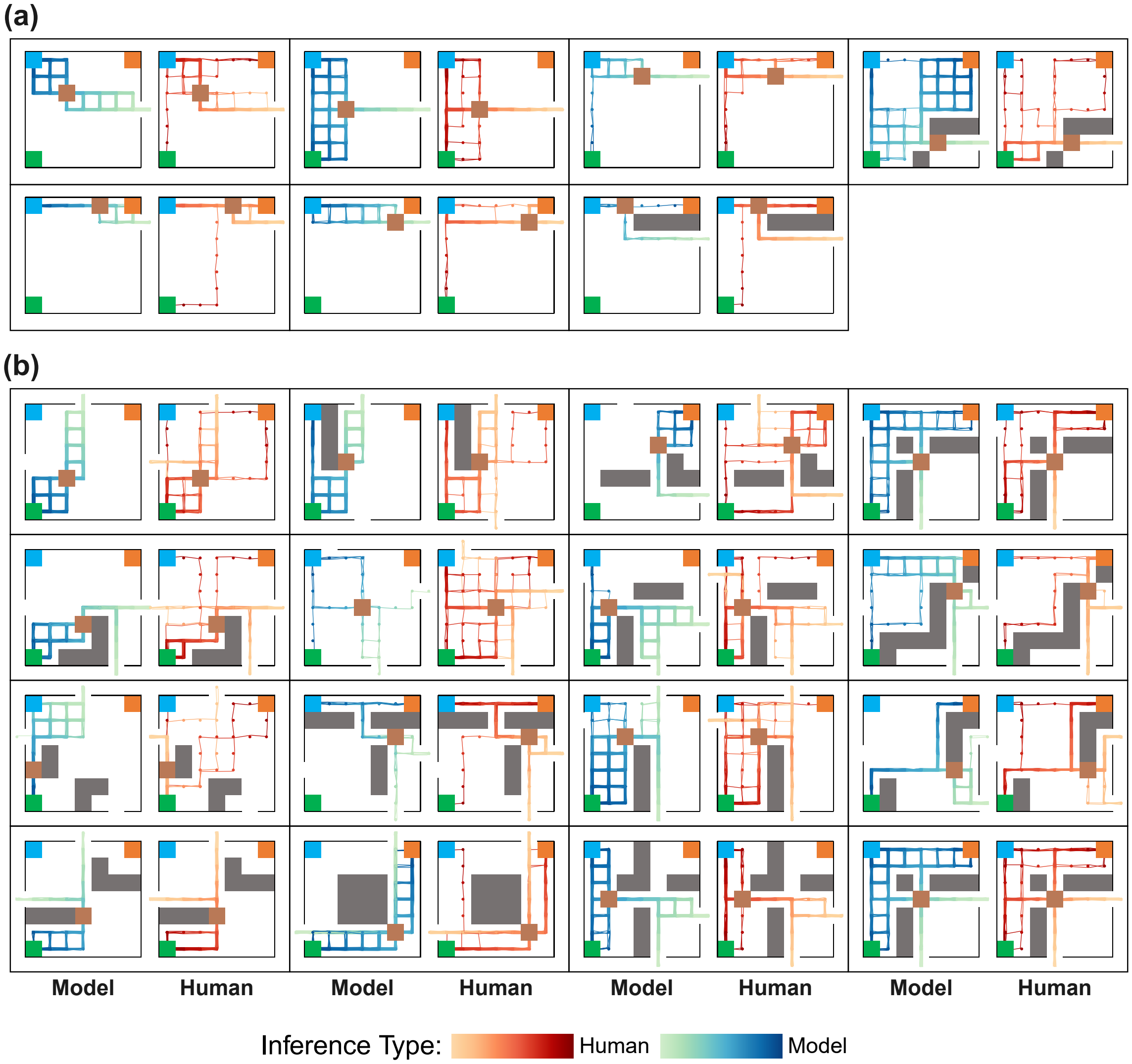

Figure 4.5: Comparison of reconstructed paths generated by our model and participants in Experiment 2. From left to right, each column of subplots corresponds to the high certainty, high negative certainty, partial certainty, and uncertain trials for goal inferences, respectively. (a) Results for trials that only had one door. (b) Results for trials that had more than one door. From top to bottom, each row of subplots corresponds to the high certainty, high negative certainty, partial certainty, and uncertain trials for door inferences, respectively. The goals A, B, and C are indicated by the blue, orange, and green squares, respectively. The doors are sequentially numbered in a clockwise order, with door 1 starting from the top (or from the right if there is no top door). Walls are marked as dark gray squares and the pile of cookie crumbs are indicated by the brown squares. Each line represents a reconstructed path, color-coded to indicate time, moving from light orange to dark red (for participants) or light green to dark blue (for the model).

generated path reconstructions relative to the baseline model ($t(39) = 9.10$, $p < 0.001$ using a Bayes factor of 1 as the reference level).

Figure 4.5 shows trial-by-trial results from Experiment 2. Each trial is presented twice, with our model's path reconstructions on the left and participant-generated path reconstructions on the right. All paths are color-coded to indicate time (with darker colors occurring later in time). For both our model and participants, the higher path density indicates where the majority inferred the agent to have traveled. As this figure shows, the distribution of participant-generated path reconstructions largely matched those generated by our model (although participants were more likely to generate suboptimal paths).

4.6 Do explicit event reconstructions in Experiment 2 predict inferences from Experiment 1?

The previous results showed that that people can not only reconstruct agents' actions, but do so in a way similar to our model. According to our proposal, this event reconstruction underlies people's capacity to infer agents' goals and entry points in Experiment 1. If this is the case, then the path reconstructions from Experiment 2 should have predictive power over the inferences that participants made in Experiment 1. To test this possibility, we extracted the goals and doors from the participant-generated path reconstructions. To achieve this, we calculated the proportion of paths that originated from each possible entrance, and the proportion of paths that reached each possible goal, and compared these values to the corresponding goal and door inferences from Experiment 1a. Figure 4.6 shows the results from this analysis. Overall, the goals and doors extracted from the participant-generated path reconstructions showed a correlation of $r = 0.89$ (95% CI: $0.83 - 0.92$) with the inferences participants made in Experiment 1a, and the strength of this fit was similar when looking only at goals ($r = 0.88$; 95% CI: $0.80 - 0.93$) or doors ($r = 0.90$; 95% CI: $0.82 - 0.95$). Furthermore, when we compared these extracted goals and doors against our model's predictions in Experiment 1a, we found a correlation of $r = 0.86$ (95% CI: $0.79 - 0.91$), and a similar fit when looking only at goals ($r = 0.85$; 95% CI: $0.76 - 0.91$) or doors ($r = 0.88$; 95% CI: $0.78 - 0.93$).

Critically, participants in Experiment 2 could only generate a single path per trial. By combining the paths of multiple participants, we were able to reveal distributions over goals and doors that quantitatively resembled the inferences participants made in Experiment 1a. The fact that these distributions predicted inferences from Experiment 1a suggests that generated paths were samples from the posterior distribution (rather than maximum likelihood or maximum *a posteriori* estimates, which would not contain enough information to reconstruct the full probability distribution over inferences). This analysis suggests that participants in Experiment 2 had access to and sampled paths in accordance to these goal and door distributions.

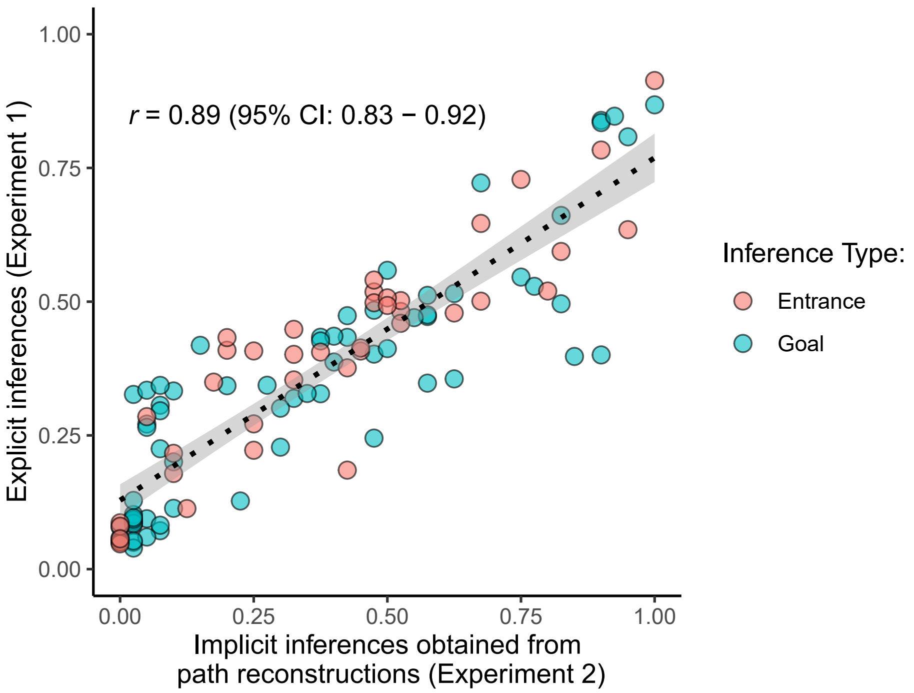

Figure 4.6: Comparison between the extracted goals and doors from Experiment 2 and the participant inferences from Experiment 1a. Color indicates inference type and the dotted line shows the best linear fit with 95% confidence bands (in light gray).

4.7 Experiment 3

Experiment 1 showed that people can infer an agent's goals and origins, and that these inferences exhibit the quantitative structure predicted by a model of event reconstruction. Experiment 2 further showed that people could explicitly reconstruct the paths in a way similar to our model. In Experiment 3, we test a further prediction of our account: If our model of event reconstruction is correct, then people should not only be able to infer a *single* agent's probable actions and goals, but also be able to estimate how many agents might have been in a room, based on how many path reconstructions are needed to explain a given scene. This study was pre-registered; all study materials can be found at `https://osf.io/q3ct5/`.

4.7.1 Participants

40 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($M = 37.62$ years, $SD = 11.94$ years).

4.7.2 Stimuli

Stimuli consisted of 15 gridworld images that were similar to those in Experiment 1, with the difference that each trial now has two piles of cookie crumbs instead of one (see Figure 4.7 for examples). Our stimuli set was designed to capture different types of inferences that our model supports. Specifically, we designed three different trials for each of the following possible inference patterns: high certainty that one agent was in the room (DEFINITELY ONE trials), partial certainty that one agent was in the room (PROBABLY ONE trials), uncertainty whether it was one or two agents in the room (UNCERTAIN trials), partial certainty that two agents were in the room (PROBABLY TWO trials), and high certainty that two agents were in the room (DEFINITELY TWO trials).

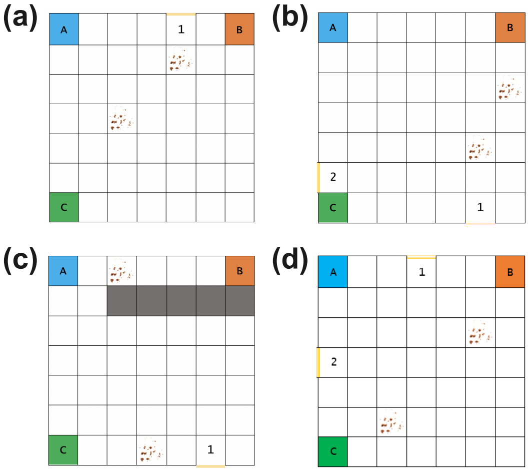

Figure 4.7: (a-d) Example stimuli from Experiment 3 for DEFINITELY ONE, PROBABLY ONE, PROBABLY TWO, and DEFINITELY TWO trials, respectively (see Experiment 3 Stimuli for details). Potential goals are positioned in the corners, labeled alphabetically, and color-coded. Doors are shown in yellow and coded numerically. Walls are shown in dark gray. Each trial included two piles of cookie crumbs positioned in various parts of the room.

4.7.3 Procedure

The procedure was nearly identical to Experiment 1a, except that participants were instead shown two piles of cookie crumbs and were told that their task was to infer if one or two agents had been in the room. After the introduction, participants completed a questionnaire that ensured they read and understood the instructions. Participants that failed at least one question were redirected to the beginning of the instructions and given a second chance to participate in the study. Participants that failed the questionnaire twice were not permitted to participate in the study.

Participants completed all 15 trials in a random order. On each trial, participants answered a multiple-choice attention-check question ("Which corner is the farthest walk from Door 1? If there is more than one correct answer, just choose one of them.") and were asked to infer how many agents were in the room ("How many people were in the room?") using a continuous slider (ranging from 0, labeled as "definitely one," to 1, labeled as "definitely two"). Participants were allowed to submit their responses for each trial only when they correctly answered the attention-check question. Otherwise, participants were told to "please pay attention and try again."

4.7.4 Model Predictions

To predict how many agents might have been in a scene we computed the probability that a agents were in scene s, through

$$p(a|s) \propto p(s|a)p(a),$$

where $p(a)$ is a prior over the number of agents that could have been present. In natural contexts, this prior should reflect the statistics of how often different agents might interact in different environments. To model our experiment, however, we used a simple uniform prior over the possibility of having one or two agents. This prior was then weighted by the likelihood of a particular number of agents a generating scene s, given by

$$p(a|s) \propto \begin{cases} \sum_{t \in \mathbb{T}} p(s|t)p(t) & a = 1 \\ \sum_{t_1, t_2 \in \mathbb{T}} p(s|t_1, t_2)p(t_1)p(t_2) & a = 2 \end{cases}$$

To compute the likelihood that two trajectories explain the scene (i.e., $p(s|t_1, t_2)$), we modified our generative model to sample two sets of entry points, goals, and trajectories at a time instead of one, where the likelihood is defined as $1/(|t_1| + |t_2|)$ if there was a scene match (i.e., both piles of cookie crumbs lie within both trajectories, and each trajectory was responsible for one of the piles) and 0 otherwise.

4.7.5 Results

Participant judgments were averaged across trials and compared against our model's predictions. Figure 4.8 shows the results from Experiment 3. Participant's relative confidence about the number of agents in the scene was quantitatively similar to our model's predictions, yielding a correlation of $r = 0.76$ (95% CI: $0.43 - 0.91$). As before, participants' pattern of data did not only qualitatively identify the best inference, but also revealed a graded pattern of confidence that is broadly consistent with event reconstruction.

Figure 4.9 shows our model's results as a function of each trial. In each subplot, the image at the top shows an abstract schematic of the trial, with both piles of cookie crumbs marked as brown squares. From left to right, each column corresponds to the DEFINITELY ONE, PROBABLY ONE, UNCERTAIN, PROBABLY TWO, and DEFINITELY TWO trials, respectively. This figure reveals how our model quantitatively predicts participant judgments across the various trials and levels of uncertainty.

Interestingly, the model fit in Experiment 3 was lower relative to Experiment 1. Under our account, this difference may arise because Experiment 3 requires reconstructing paths for a single agent, reconstructing paths for multiple agents, and

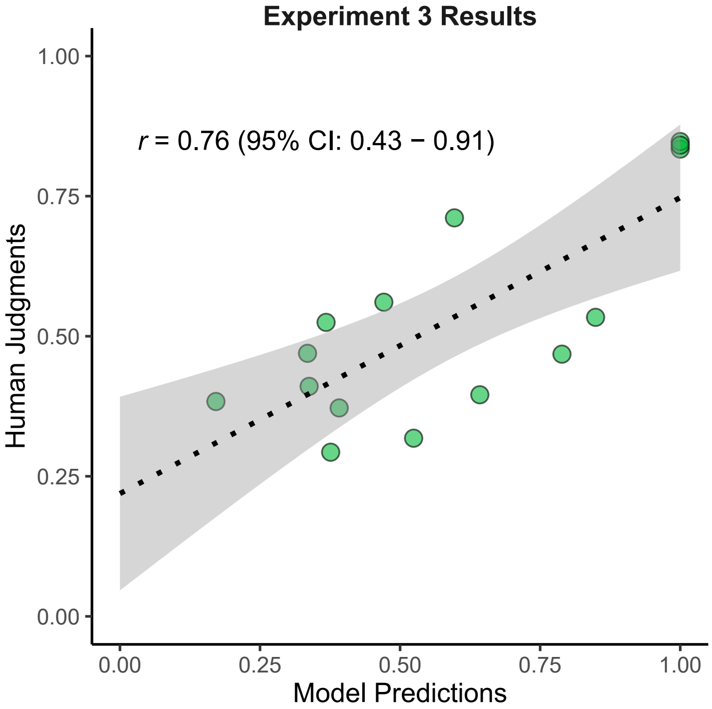

Figure 4.8: Results from Experiment 3. Each point corresponds to a judgment, with model predictions on the x-axis and mean participant judgments on the y-axis. The dotted line shows the best linear fit with 95% confidence bands (in light gray).

weighting their relative probability of generating the observed scene. Consistent with this, we found higher mismatches between our model and participants in the PROBABLY trials ($MSE = 0.053$) over the DEFINITELY ($MSE = 0.021$) and UNCERTAIN trials ($MSE = 0.019$). That is, participants struggled more in trials that relied on a capacity to make precise comparisons between the number of single-agent reconstructions and two-agent reconstructions.

As in Experiment 1a, we also evaluated whether participant judgments could be explained by superficial features of the stimuli rather than via event reconstruction. We tested this possibility through a logistic regression trained to predict participants' distribution over the number of agents they thought were in the room as a function of the distance between each goal and each pile of cookie crumbs, the average distance between each pile of cookie crumbs and the doors, the number of doors, and all of their interactions. We trained and tested this alternative model in the same way as the one described in Experiment 1a.

Even though this alternative model had access to the qualitative structure of participant judgments, it nonetheless produced a correlation of $r = 0.19$ (95% CI: $-0.30 - 0.66$) with participant judgments, which was substantially lower than the one produced by our model ($\Delta r = 0.58$; 95% CI: $0.12 - 1.17$). These results extend our findings from Experiments 1 and 2, suggesting that people can not only infer an agent's goals and origins based on indirect evidence of their presence, but also

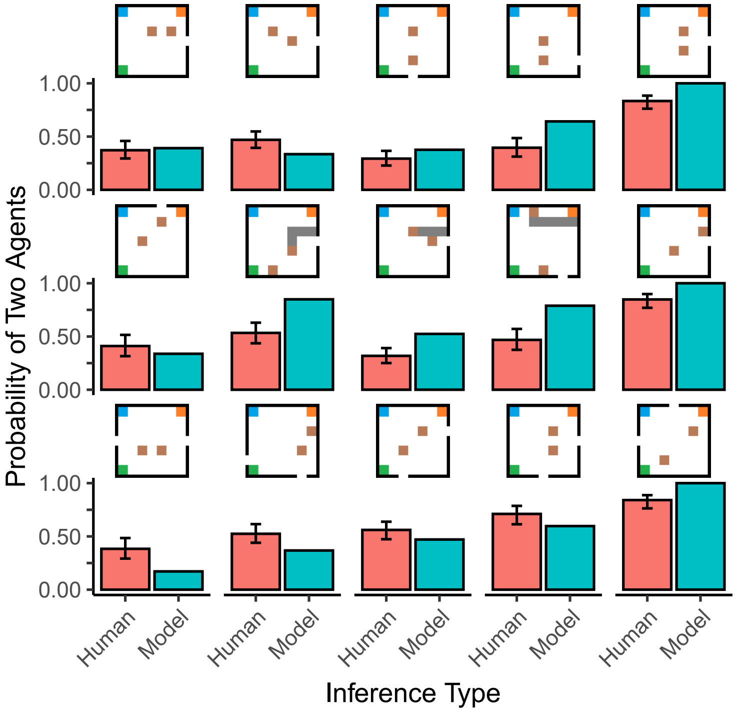

Figure 4.9: Detailed results from Experiment 3. From left to right, each column corresponds to DEFINITELY ONE, PROBABLY ONE, UNCERTAIN, PROBABLY TWO, and DEFINITELY TWO trials, respectively. Red bars represent mean participant judgments and blue bars represent our model's predictions. Error bars on participant judgments represent 95% bootstrapped confidence intervals.

whether multiple agents may have been present in a given scene.

4.8 Discussion

Research on human action understanding has historically focused on how we infer the goals and mental states of agents whose behavior we are observing. Our results show that our capacity to reason about others goes beyond face-to-face interactions and includes nuanced social inferences from simple physical scenes. In Experiment 1, we showed that people can infer an agent's goals (i.e., where an agent was going) and past actions (i.e., where an agent came from) from a single piece of indirect evidence about their presence. The tight correspondence between our model's predictions and the fine-grained structure of participant judgments suggested that these inferences were structured around a form of mental event reconstruction: people infer the actions that an agent took, and use this reconstructed behavior to make richer social inferences. Experiment 2 showed further support for our proposal, revealing that people can explicitly reconstruct the actions that someone took based on indirect physical evidence, in a way similar to our model. Furthermore, these explicit reconstructions predicted participant inferences in Experiment 1, showing a direct link between people's ability to reconstruct behavior from physical evidence, and the corresponding social inferences that they make. Finally, in Experiment 3, we found that people can also infer how many agents were in a given scene, based on the number of paths they needed to reconstruct to explain the scene.

4.8.1 What cognitive capacities are required for event reconstruction?

Our computational model formalized social inferences as the process of reconstructing behaviors that explain the observed physical evidence. Our model's quantitative fit with participant judgments, and the failure of our alternative models (despite being trained on participant judgments), suggests that people were performing similar computations. In particular, the similarity between the paths generated by our model and those drawn by participants (see Figure 4.5) suggests that social inferences from physical evidence are tied to a form of event reconstruction.

The heart of our proposal—expressed in Equation 4.2 (see Section 4.2)—posits that event reconstruction depends on two different cognitive capacities. The first is a model of how agents act in the world. The second is a model of how agents' actions leave observable traces in the environment.

In our model, the first capacity consisted of a simple expectation that agents navigate towards their goals as efficiently as possible, given the environmental constraints. This expectation, known as a *teleological stance* (Gergely, 2003; Gergely and Csibra, 1997), has been hypothesized to be a precursor to mental-state reasoning, supporting simple social inferences without requiring active representations of other people's minds (Gergely and Csibra, 2003). From this standpoint, our computational model shows that a full-fledged Theory of Mind is not necessary for performing social reconstructions from physical evidence, and a teleological stance can suffice.

At the same time, agents with a Theory of Mind might be able to derive richer social inferences. To illustrate this, imagine that a valuable object that was hidden in a closet in someone's house has gone missing. Suppose also that drawers and cabinets throughout the house were left open, but nothing else had been taken. In this situation, a pure teleological stance could reveal that the thieves navigated through the house opening drawers and cabinets. However, a teleological stance alone would end there, failing to reveal *why* the thieves pursued these goals. This event, analyzed through a Theory of Mind, however, would reveal that the thieves knew that the valuable object was in the house, did not know its exact location, and therefore searched the house to find it.

This example raises the possibility that a non-mentalistic teleological stance enables people to reconstruct the actions that an agent took, by assuming that they navigate efficiently in space. Once these actions have been reconstructed, our Theory of Mind might enable us to extract the complex mental states that can explain why the agent took the actions that they did. This is a direction that we hope to explore in future work.

The second capacity implemented in our model is an understanding of how actions leave observable traces in the environment. Our model therefore posits that event reconstruction requires an ability to associate different actions with their corresponding observable traces. Our model used a highly simplified setting where the observable evidence consisted of a small pile of cookie crumbs. In more realistic situations, the types of traces that agents leave behind can be rich and variable, from unambiguous cues like foot tracks on the ground, to more subtle ones, like finding a single apple tree with no apples, in a row of trees full of ripe apples. This suggests that people's capacity to reconstruct behavior is simultaneously powered and constrained by their knowledge of the relationship between actions and physical traces.

While our work focused on adults, some recent research suggests that these capacities might emerge in early childhood. In particular, preschoolers can judge what types of physical constructions (such as different types of block towers) require more physical effort (Gweon et al., 2017), suggesting an early understanding between actions and physical outcomes. Moreover, children can also determine what actions are more likely to leave physical traces. For example, lifting an upside-down cup filled with rice will likely leave visible rice grains after the cup has been repositioned. But it is possible to lift and reposition an upside-down cup filled with a few large rocks without leaving any evidence behind (Jacobs et al., 2021). Recent research has found that children can even associate physical outcomes with the corresponding mental states of the agent who generated them (Pelz et al., 2020). Finally, and most strikingly, young children can infer the transfer of ideas by seeing how different agents create artifacts (Pesowski et al., 2020), a capacity known as "intuitive archaeology" (Hurwitz et al., 2019; Schachner et al., 2018). While these results point towards an early understanding of the relation between the social and physical world, to our knowledge, it is an open question whether these inferences are also linked to some form of explicit or implicit event reconstruction.

Finally, at the highest level, our work builds on the idea that human cognition is structured around mental models (also called intuitive theories) of the world (Tenenbaum et al., 2011), including intuitive theories of the physical world (Battaglia et al., 2013) and of others (Jara-Ettinger et al., 2020a). Following this tradition, our model posits that people have (i) a causal understanding of how goals lead to actions and how actions leave observable traces, and (ii) a mechanism for inverting this causal model, enabling people to move from observed traces to the underlying goals. In our model, the inversion mechanism was implemented as Bayesian inference via Monte Carlo simulations. This approach is consistent with growing evidence that action-understanding involves some form of Bayesian inference (Baker et al., 2017; Ullman et al., 2009; Jara-Ettinger et al., 2020a). Nonetheless, our work only tested our model at Marr's computational level of analysis (Marr, 1982), and it does not imply that people are specifically using a Monte Carlo based approach to implement Bayesian reasoning. Indeed, related work has found that this type of inference can be approximated via simpler strategies (Bonawitz et al., 2014), and people's inferences in our task might not have required active sampling in participants. At the same time, work in intuitive physics has found some evidence of active sampling in physical reasoning, opening the possibility that this extends to social reasoning as well (Hamrick et al., 2015). These are questions that we also hope to explore in future work.

4.8.2 Study limitations

Our work has three main limitations. First, our model and experiments focused on highly simplified events. In more realistic situations, the space of goals that an agent might pursue, and the physical evidence they leave behind is substantially more complex than what our two-dimensional gridworlds can capture. To reason about a chewed-up pencil, for example, our model would require a more extensive description of human behavior to compute how an anxious mental state shapes an agent's action space, and how the resulting candidate actions (e.g., chewing) leave traces in the environment. Our proposed model does not currently support social inferences at this level of complexity, and it is an empirical question whether our approach could capture human reasoning in these more naturalistic events.

One way in which our framework could tackle richer inferences is by using a full-fledged model of intuitive physics to evaluate how actions leave traces in the environment. A recent body of work in cognitive science has found that human intuitive physics is instantiated as a *physics engine* that supports rich probabilistic simulations of how objects and forces interact in the environment (Fischer et al., 2016; Battaglia et al., 2013), and that physical simulations might underlie how we reason about the interaction between agents and objects (Yildirim et al., 2019). Thus, using a physics engine to simulate how the forces that agents apply to the world leave observable traces might enable our computational framework to handle more complex physical events that contain social information.

Our second main limitation lies in the narrow range of inferences that we asked people to make: inferences about where an agent was going, where they entered from, and how many agents were involved. As noted above, all of these inferences can be explained through a *teleological stance* (Gergely and Csibra, 2003). Consequently, our work does not test the extent to which people can infer complex mental states or personality traits from physical evidence. Recent work has found that people can indeed make rich communicative inferences from physical arrangements of objects (Lopez-Brau and Jara-Ettinger, 2020; Sarin et al., 2021); however, in this work, the position of the objects unambiguously revealed the agent's actions (they positioned the objects where they were most visible to others). This work therefore leaves open whether the capacity to infer these types of mental states extends to events where people must perform more complex forms of event reconstruction. In future work, we hope to incorporate richer models of mental-state inference to test people's capacity to infer mental states such as beliefs, desires, knowledge, and intentions from physical evidence (Jara-Ettinger et al., 2020a; Baker et al., 2017).

Our third limitation is that our work used simple events with minimal social context: participants had nearly no information about the agent, and the goals consisted of simple abstract squares. This enabled us to test people's capacity to reconstruct events in a controlled manner. In more naturalistic situations, however, the content of the goals often reveals important information that can help people build more nuanced inferences. Imagine, for instance, that one of the squares in our stimuli was a work desk, the second one was a stationary bicycle, and the third one was a TV. With this context, the physical trace would not only allow people to infer the agent's goal, but also richer aspects of their personality. Relatedly, when more context is available, people also rely on inferred stereotypes to attribute dispositions (Gosling et al., 2002, 2008). These richer context-based inferences were not captured by our work, and are a critical challenge towards building computational models that fully capture human social reasoning.

Our work also leaves a critical question open. Our experiments focused on situations where people were explicitly told that an agent was previously present. Our work therefore does not speak to how people use physical information to infer that an agent was present in the first place. One possibility is that people engage in a pervasive and constant social analysis of all physical scenes. Doing so, however, might be prohibitively costly and unnecessary. As such, it is likely that people are attuned to the physical signatures that reveal the presence of an agent, which then trigger social reasoning from physical evidence. Consistent with this second view, research suggests that people can infer the presence of an agent based on apparent order (Newman et al., 2010; Keil and Newman, 2015b) and on a sensitivity to human-like errors that people leave behind when interacting with the world (Lopez-Brau et al., 2021). An open question is how the ability to detect the presence of an agent interacts with the ability to reconstruct their behavior and infer their mental states.

4.8.3 Implications and conclusions

At first glance, our computational framework appears to suggest that any creature with some form of naïve psychology and naïve physics ought to be able to perform social inferences from physical evidence (i.e., access to the two key components of Equation 4.2). This may not be the case, however, because our model also requires an ability to transfer information across these intuitive theories (reconstructing behavior via naïve psychology and evaluating how they compare to the environment via naïve physics). While this is an open empirical question, research suggest that intuitive physics and intuitive psychology rely on separate neural circuitry (Fischer et al., 2016; Saxe and Powell, 2006), leaving open the question of how these two intuitive theories might work in tandem to reconstruct other people's behavior from physical evidence.

One interesting case that suggests such a feat might not be simple comes from research with vervet monkeys. Vervet monkeys have an astonishing degree of social intelligence, including a nuanced repertoire of vocal calls to signal different types of predators, each associated with different escape responses (Seyfarth et al., 1980a,b). Yet, vervet monkeys routinely fail to identify predators from indirect physical evidence. For instance, vervet monkeys fail to infer that a python is hiding in a nearby bush when they encounter the distinct tracks that they leave behind. Similarly, vervet monkeys also fail to infer the presence of a leopard upon encountering a gazelle carcass on a tree (where leopards usually drag their prey so they can feed in solitude; Cheney and Seyfarth, 1985). Critically, this failure appears to persist even after vervet monkeys have, in past events, seen the direct association between the physical evidence and the predator (Cheney and Seyfarth, 1985, 2008). These results might point to the possibility that the form of event reconstruction that we present here might require capacities that go beyond simple physical and social reasoning, as they involve an ability to combine the two capacities to derive richer inferences than would be otherwise possible.

Overall, our results illustrate the sophistication of human social intelligence. Beyond being able to make social inferences about agents that we are personally interacting with, we can also make social inferences about agents we have never encountered, just from minimal indirect evidence that reveals their presence. Researchers have long argued that humans are unique in their ability to reason about and navigate the social world (Herrmann et al., 2007). Our work shows that this ability is not confined to social interactions, but can fundamentally affect how we reason about the physical world, allowing us to see social meaning embedded in physical structures, like a pile of rocks, where other animals may see merely just that: a pile of rocks.

Chapter 5

How do we infer the communicative meaning of objects?

This chapter is based on: Lopez-Brau, M. & Jara-Ettinger, J. (*under review*). People can use the placement of objects to infer communicative goals.

Abstract

Beyond words and gestures, people have a remarkable capacity to communicate indirectly through everyday objects: A hat on a chair can mean it is occupied, rope hanging across an entrance can mean we should not cross, and objects placed in a closed box can imply they are not ours to take. How do people generate and interpret the communicative meaning of objects? We hypothesized that this capacity is supported by social goal inference, where observers recover what social goal explains an object being placed in a particular location. To test this idea, we study a category of common ad-hoc communicative objects where a small cost is used to signal avoidance. Using computational modeling, we first show that goal inference from indirect physical evidence can give rise to the ability to use object placement to communicate. We then show that people from the U.S. and the Tsimane'—a farming-foraging group native to the Bolivian Amazon—can infer the communicative meaning of object placement in the absence of a pre-existing convention, and that people's inferences are quantitatively predicted by our model. Finally, we show evidence that people can store and retrieve this meaning for use in subsequent encounters, revealing a potential mechanism for how ad-hoc communicative objects become quickly conventionalized. Our model helps shed light on how humans use their ability to interpret other people's behavior to embed social meaning into the physical world.

5.1 Introduction

Humans have a remarkable capacity to communicate through objects, even ones we do not usually think of as conveying meaning. A hat on a chair can reveal that the seat is taken; rope surrounding a patch of grass can tell us not to walk through; and, during snowy winters in the northeastern United States, plastic chairs on shoveled parking spots are used to signal that they are not up for grabs. These kinds of everyday objects (Figure 5.1) do little to physically constrain our actions, yet they affect our behavior because we recognize the meaning they convey. Consistent with this, past empirical research has shown that people spontaneously use objects to communicate (e.g., leaving an open notebook on a library table to mark that the space is occupied; Sommer and Becker, 1969; Becker and Mayo, 1971; Edney and Jordan-Edney, 1974), and detect when an object is communicative (e.g., realizing that the table with a notebook must be taken; Becker, 1973; Shaffer and Sadowski, 1975), with this ability possibly emerging in childhood (Rossano et al., 2015).

What are the cognitive capacities that support our ability to communicate through objects? One possibility is that communicative objects emerge from a system of simple conventions, where objects and their placement are explicitly associated with different communicative meanings. As children, for instance, most of us likely ignored strap barriers at banks, movie theaters, and DMVs, and their meaning had to be explicitly taught to us. After learning their meaning, we were then able to recall it whenever we encountered them in new locations.

While conventional knowledge is undoubtedly a major driver for how we learn and use communicative objects, people are also able to generate novel communicative objects that others can readily understand (such as placing an ironing board to mark that someone has reserved a parking spot; Figure 5.1i). What computations underlie this capacity? And how does the communicative meaning of novel objects become conventionalized?

Here we hypothesized that the capacity to embed and infer communicative meaning from novel objects emerges from our ability to reason about the mental states behind other people's behavior—our *Theory of Mind* (ToM; Wellman, 2014; Gopnik et al., 1997). The central idea in our proposal is that, if people can infer other agents' mental states based on how they manipulated an object (via Theory of Mind), then people can also strategically manipulate objects with the purpose of eliciting mental-state inferences in agents who encounter these objects. Through this method, people can intentionally manipulate their environment with the goal of communicating their desires to people who navigate the environment when the communicator is absent. We propose that this type of reasoning might support the creation of novel ad-hoc communicative objects, which can then quickly become conventionalized and widespread, supported via memory and recognition.

To explore this idea, this paper focuses on a family of objects like those shown in Figure 5.1. These objects are often not intrinsically communicative: Hats, chairs, and rope are not purposefully designed for communication, but they can nonetheless convey a message when placed in certain locations (Figures 5.1a-c). Moreover, despite their varied use, these objects all communicate some kind of restriction (e.g., “do not use” or “do not cross”). Critically, however, this restriction is not imposed purely through a physical constraint: The cost that these objects impose on agents is low enough that it could be easily ignored (e.g., walking over the “barriers” in Figures 5.1b and 5.1k is trivial). Intuitively, these objects instead work because people realize that the object was intentionally placed with the purpose to communicate. Because of this common structure, we will refer to these objects as *low-cost communicative blockers* (LCCBs). While these objects do not capture the full scope of everyday communicative objects, we believe their use is a fruitful case study for understanding our proposal. We return our focus to communicative objects more broadly in the Discussion.

To illustrate the logic of our proposal, imagine trying to find the exit of an unfamiliar building. As you walk down a hallway, you find two doors, side by side. Suppose, however, that one of the doors has a broom positioned diagonally across it. Naturally, it is easy to recognize that (1) someone intentionally placed the broom there and that (2) it creates a small inconvenience for people wanting to walk through the door. When considering why someone would choose to use a broom to block a door, one possibility is that they wanted to prevent people from walking through. But if that were the case, why not put more effort into blocking the door, given how easy it is to move the broom out of the way? Intuitively, this is because their goal was not to create an insurmountable physical constraint—which would require more effort to achieve—but rather to prompt you to infer that they do not want you to walk through.

This proposal assumes that people can detect intentional arrangements of objects (e.g., a broom placed diagonally across a door was likely placed intentionally), infer what an agent did (e.g., an agent must have taken the broom and placed it there), and determine how much effort it required from the agent and how much it affects us (e.g., how hard was it to place the broom and what effects does this have on my potential plans?). Consistent with this, past research has shown that people have a rich understanding of what physical environments reveal about people (Gosling et al., 2002; Hurwitz and Schachner, 2020). Moreover, people can infer others’ actions from indirect physical evidence of their presence (Lopez-Brau et al., 2022), and estimate the effort involved in moving and manipulating objects (Yildirim et al., 2019). These capacities also emerge early in development, with children drawing surprisingly rich inferences from physical evidence, ranging from inferences about what actions an agent took (Jacobs et al., 2021) and what they knew (Pelz et al., 2020) to inferences about even richer social information, such as whether two people transmitted ideas (Pesowski et al., 2020) and have shared interests (Pesowski et al., 2021).

Critically, for communicative objects to have their intended effect, the ability to reason about them is not enough: people must also be motivated to behave cooperatively. If this were not the case, people would ignore low-cost communicative blockers

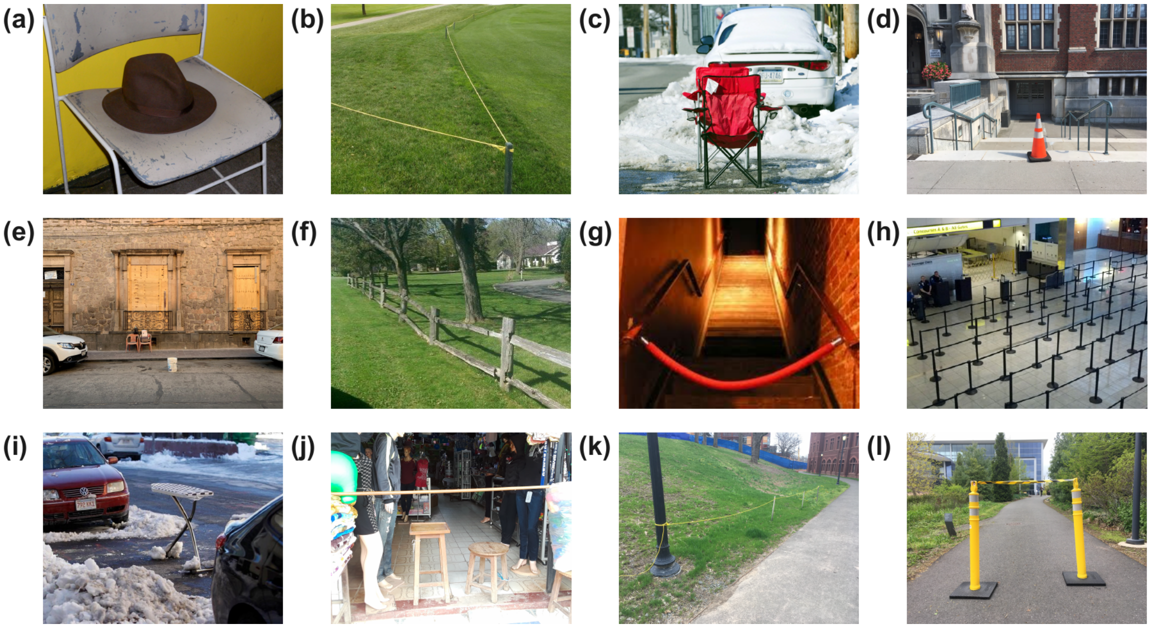

Figure 5.1: Real-world examples of people communicating through objects. (a) A hat on a chair indicating that someone intends to return. (b) Rope a few inches above the grass so that people know not to walk through. (c) Chairs along the side of the street in South Boston to reveal that someone shoveled and claimed this parking spot. (d) A traffic cone in front of some stairs signaling limited access. (e) A bucket along the side of the street in central Mexico indicating that the parking spot is reserved. (f) An easy-to-cross fence marking a property limit. (g) A stanchion across a stairwell revealing access may be restricted to certain individuals. (h) Belt barriers at the airport telling passengers that they should form a line (and where). (i) An ironing board along the side of the street indicating that the parking spot is taken. (j) A wooden pole and two small benches in a store in Bolivia indicating that the owner is not available. (k) A small rope along a sidewalk asking people not to walk near a construction site. (l) A pair of traffic posts deterring people from using this walkway.

(LCCBs; since the cost they impose is negligible), and communicators would favor creating insurmountable physical constraints rather than communicative signals. While there are undoubtedly cases where people ignore LCCBs, and where people build physical barriers because they do not expect cooperativeness, the pervasive use of these objects suggests that there are many cases where people expect strangers to cooperate by default. This is consistent with evidence that even young children will spontaneously cooperate with strangers (Warneken and Tomasello, 2006) and that adults have a default propensity to cooperate (Rand, 2016).

While all this past work establishes the cognitive pre-requisites that our proposal builds on, to our knowledge, no work has yet explored specifically whether these capacities underlie the ability to use communicative objects (Figure 5.1).

5.1.1 Paper overview

Our paper has three goals. Our first goal is to test whether our theoretical proposal can, in principle, explain the logic of low-cost communicative blockers (LCCBs), where agents share mental states by using objects to impose a minimal cost on observers. To achieve this, we present a model that explicitly formalizes our proposal in computational terms, and we explore its behavior in synthetic simulations (Sections 5.2-5.3) to test whether it can produce patterns that resemble how people use communicative objects. Our computational model focuses on the inferences that we make once an object placement is detected as intentional. We return to the question of how to detect intentional placements in the Discussion.

Having found theoretical support for our proposal, our second goal is to test whether the fine-grained quantitative predictions of our account match human judgments when reasoning about novel low-cost communicative blockers. That is, we use our model to generate exact numerical predictions about the strength of inferences that people should make in different situations, and we compare them to human judgments (Experiment 1).

Finally, our third goal is to test whether the mechanisms we propose play a role when conventional knowledge (i.e., object-meaning mappings that are in common ground for a social group) is unavailable. That is, our goal is not to argue against the critical role of convention, but rather to ask what types of inferences people engage in when they face an object that has no conventional meaning attached to it (Experiments 2-3), and to explore how these inferences become conventionalized (Experiment 4).

While our account proposes mechanisms that support both the creation and understanding of communicative objects, these two behaviors are asymmetrical in two ways. First, as we show in our model below, recognizing the meaning of communicative objects is easier than creating them, requiring one fewer level of recursion. Second, for communicative objects to become ubiquitous, the ability to infer their meaning must be widespread, while the ability to invent them can be restricted to a few individuals. Thus, after confirming the computational plausibility of our account (Sections 5.2-5.3), our behavioral studies (Experiments 1-4) focus on people's ability to infer the communicative meaning of low-cost communicative blockers, rather than on how they are created. We return to this asymmetry in cognitive demand in the Discussion.

5.2 Computational Framework

For simplicity and clarity, we describe our model in the context of a simple event similar to the ones we use in Experiments 2-4. Here, an agent (the *decider*) encounters two doors—door A and door B—and must decide which one to walk through. Before they do, another agent (the *enforcer*), who wants to influence the decider's choice, has the opportunity to place objects in front of either door, including stacking multiple objects to create a physical constraint. To illustrate, Figure 5.2a shows a situation where the enforcer has access to four boulders that can be positioned in front of either door. Figures 5.2d-h show five possible changes that the enforcer could implement (among many others). Our model therefore consists of (i) an enforcer that moves objects in a scene with the goal of affecting a decider's behavior, and (ii) a decider that determines what to do in a scene by thinking about the costs that the objects in the scene impose. Critically, we assume that the enforcer and decider are never in the scene at the same time, such that the decider only has access to the physical layout of the scene.

In this setup, the enforcer can always pursue a simple non-communicative strategy: stack enough objects in front of one of the doors to the point that walking through it is so much work that the decider will prefer to avoid it (e.g., Figure 5.2f). However, stacking objects in front of a door is also costly for the enforcer. This creates a preference for more efficient strategies, where agents might exploit their Theory of Mind to use objects in a communicative manner.

Under our proposal, people use objects to share mental states by reasoning about the costs incurred by the enforcer (how costly is it for the enforcer to block paths?) and the decider (what costs does this impose on the decider?). Because past work has already studied how people reconstruct behavior from physical displays and estimate the underlying costs (Yildirim et al., 2019; Lopez-Brau et al., 2022; Pesowski et al., 2020), our model takes this capacity for granted and focuses on the inferences that people make given access to these costs.

To make cost-based inferences, our framework instantiates Theory of Mind (ToM) as a form of simple recursive social reasoning, similar to models developed to understand pedagogical demonstrations (Ho et al., 2016; Shafto et al., 2014), pragmatics (Frank and Goodman, 2012; Goodman and Frank, 2016), and mental-state inferences (Ullman et al., 2009), and similar to the logic behind k-level ToM models (where k is a variable indicating the recursion depth within ToM; Devaine et al., 2014). At its core, our model is structured around an assumption that agents act to maximize their subjective utilities—the difference between the costs that they incur and the rewards that they obtain. This assumption is at the heart of human mental-state inferences in adults (Jern et al., 2017; Baker et al., 2017; Jara-Ettinger et al., 2020b) and emerges early in development (Liu et al., 2017; Jara-Ettinger et al., 2016; Gergely and Csibra, 2003; Lucas et al., 2014).

Formally, let S be the space of all possible scenes, where each scene $s \in S$ represents an observable arrangement of objects (e.g., see Figures 5.2a,d-h for six possible scenes in the boulder example). Each agent in this context is defined by two main components. The first is a cost function that captures how agents interact with the environment. For the enforcer, their cost function C_E represents the cost of moving objects, such that $C_E(s_0, s)$ is the cost of transforming an initial scene s_0 into a final scene s (e.g., the cost of changing the scene so that an object in a corner is now in front of a door). For the decider, their cost function C_D represents the cost of navigating the environment, such that $C_D(a, s)$ is the cost of taking action a in scene s (e.g., the cost of walking through a door with an object in the way).

The second main component is a reward function that captures each agent's desires. The enforcer's reward function R_E represents their desire to affect the decider's behavior. That is, $R_E(a)$ is the reward the enforcer obtains when the decider takes action a. The decider's reward function R_D represents their own personal desires: $R_D(a)$ is the decider's personal reward when choosing action a (e.g., the decider's reward when they choose door A in Figure 5.2).

Our computational framework uses these cost and reward functions to build a model of recursive social reasoning, where the enforcer decides how to move objects by thinking about what action they hope the decider will take, and the decider decides what action to take by inferring what the enforcer wants, based on how they manipulated the objects in the scene. Below, we present the logic of our model, starting with the grounding level of the recursive structure. A more detailed presentation of our model can be found in SM.

5.2.1 Non-mentalistic decider

The lowest level of our model consists of a non-mentalistic decider D_0 that represents an agent lacking any awareness that objects in a scene may have been intentionally manipulated by another agent. This decider therefore chooses what to do based on the physical properties of the scene alone. Given a scene s, the non-mentalistic decider's utility for taking action a is given by:

$$U_{D_0}(a; s) = R_{D_0}(a) - C_{D_0}(a, s),$$

where $R_{D_0}(a)$ is the reward that the decider obtains from taking action a and $C_{D_0}(a, s)$ is the cost they incur from taking that action in scene s.

We transform this utility function into a probability distribution over actions by applying the *softmax function*:

$$p_{D_0}(a|s) \propto \exp(U_{D_0}(a; s)/\tau).$$

The softmax function is a standard method for transforming utility functions into probability distributions, guided by a temperature parameter $\tau \in (0, \infty)$. When τ is low, the decider consistently chooses the actions that maximize the utility function (converging towards optimal behavior as $\tau \to 0$). When τ is high, the decider's behavior becomes noisier, and the agent is more likely to select actions that are not necessarily the best ones (converging towards random behavior as $\tau \to \infty$).

5.2.2 Simple enforcer

The next level of our model consists of a simple enforcer E_0 who reasons about the non-mentalistic decider D_0. That is, this enforcer determines what to do under the assumption that the decider will not realize that the objects contain any social information and will instead see them as nothing more than physical obstacles.

Formally, suppose the world is in some initial state s_0 and the enforcer wants the decider to take action a (e.g., the initial scene might be Figure 5.2a, and the enforcer wants the decider to choose door A). To do this, the simple enforcer considers different possible scenes s (e.g., Figures 5.2d-h) and evaluates them through the utility function:

$$U_{E_0}(s; a, s_0) = \underbrace{R_{E_0}(a) p_{D_0}(a|s)}_{\substack{\text{Expected reward} \\ \text{when decider takes} \\ \text{action } a \text{ in scene } s}} - \overbrace{C_{E_0}(s_0, s)}^{\substack{\text{Cost of transforming} \\ \text{scene } s_0 \text{ into scene } s}}.$$

Here, the first term ($R_{E_0}(a) p_{D_0}(a|s)$) is the enforcer's expected reward (i.e., the reward they obtain when the decider takes action a, weighted by the probability that the decider takes this action). This enforcer's ability to predict how the decider will act, $p_{D_0}(a|s)$, is computed using the non-mentalistic decider model (Equation 5.2.1, Section 5.2.1). This term is then balanced against the cost $C_{E_0}(s_0, s)$ that the enforcer incurs in transforming scene s_0 into scene s. Combined, the first term leads the enforcer to prefer scenes where the decider is more likely to take the desired action, and the second term leads the enforcer to favor minimal scene changes over drastic ones.

5.2.3 Mentalistic decider

Having defined the simple enforcer E_0, we can now specify a mentalistic decider D_1 that reasons about this enforcer's choices. That is, this decider infers why the enforcer decided to modify the scene, and takes this into account when deciding what to do.

Formally, the mentalistic decider assigns a utility to each action via

$$U_{D_1}(a; s_0, s, \phi) = \underbrace{R_{D_1}(a) - C_{D_1}(a, s)}_{\substack{\text{Decider's egocentric} \\ \text{costs and rewards}}} + \overbrace{\phi < \ell(a|s_0, s) >}^{\substack{\text{Decider's allocentric} \\ \text{preferences}}}.$$

The first two terms ($R_{D_1}(a)$ and $C_{D_1}(a, s)$) capture the decider's egocentric rewards and costs for taking action a in scene s, respectively (identical to the utility function for the non-mentalistic decider; Equation 5.2.1, Section 5.2.1).

The final term, $\phi < \ell(a|s_0, s) >$, represents the decider's utility for acting in accordance with the enforcer's preferences, also known as the decider's "adopted utility" (Powell, 2022). Here, $< \ell(a|s_0, s) >$ is the decider's belief that the enforcer wants them to take action a, based on the change from scene s_0 to scene s. This term is then weighted by a real-valued cooperation parameter ϕ that captures the decider's motivation to pursue, act against, or ignore the enforcer's preferences. When ϕ is positive, the decider is motivated to act in a way that is consistent with the enforcer's preferences. Conversely, when ϕ is negative, the decider is antagonistic and prefers to act against the enforcer's preferences. Finally, when $\phi = 0$, the decider acts egocentrically (becoming the same model as the non-mentalistic decider), and treats objects as physical constraints, ignoring why the enforcer might have positioned them there.

Critically, the decider does not know a priori what the enforcer wants them to do (i.e., the decider does not have direct access to $\ell(a|s_0, s)$). This term is therefore inferred by considering the enforcer's possible reward functions:

$$< \ell(a|s_0, s) >= \int_{R_{E_0} \in R} \mathbb{1}_{\arg\max R_{E_0}}(a) p(R_{E_0}|s_0, s).$$

This equation adds up the probability of every possible reward function $R_{E_0} \in R$ where a is the preferred action. For each of these reward functions, its probability, $p(R_{E_0}|s_0, s)$, is inferred by reasoning about the enforcer's choice to change scene s_0 into scene s:

$$p(R_{E_0}|s_0, s) \propto p_{E_0}(s|R_{E_0}, s_0)p(R_{E_0}),$$

with the likelihood $p_{E_0}(s|R_{E_0}, s_0)$ computed using the simple enforcer model (i.e., it is given by the softmax of Equation 5.2.2, Section 5.2.2).

Note that this formulation assumes that the mentalistic decider knows both the scene's initial and final states (s_0 and s). This allows deciders to infer the enforcer's rewards by reasoning about the costs that were introduced. In more realistic situations, it is more likely that deciders have a prior distribution over scenes ($p(s_0)$) rather than perfect knowledge about the initial scene s_0. Returning to the example in the introduction, for instance, when encountering a broom placed across an entrance, a decider may not know where the broom was situated before an enforcer placed it across the door, but they may believe it was more likely that it was positioned elsewhere. Modeling prior expectations about scene distributions is beyond the scope of our model, but we return to the implications of this assumption in the Discussion.

5.2.4 Complex enforcer

Finally, we can define a complex enforcer E_1 that modifies scenes by thinking about a mentalistic decider D_1. This model is identical to the simple enforcer (Section 5.2.2), with the only difference that it predicts the decider's behavior using the mentalistic decider model (Section 5.2.3), rather than the non-mentalistic one. That is, the term $p_{D_0}(a|s)$ from Equation 5.2.1 is now replaced by $p_{D_1}(a|s, \phi)$ (i.e., the softmax of the utility function in Equation 5.2.3). This enforcer can therefore manipulate scenes under the assumption that the decider will attempt to decode their preferences.

5.2.5 Model implementation details

The computational framework specified above captures the proposal that people extract costs from physical scenes (i.e., what is the cost of taking different actions, and what costs did another agent incur in positioning the objects), and use them to make mental-state inferences. Because past work has already studied how people might infer costs from physical scenes (Yildirim et al., 2019), our interest is in testing how cost manipulations shape communication with objects. Therefore, in our model, we directly provide the costs associated with each scene change, which enables us to focus on the contribution of cost-based reasoning.

In principle, the parameters in our model can all be real-valued. For simplicity, we bounded costs and rewards to integers in the range 0 to 9. This enables us to easily interpret the range with 0 being null costs and rewards, and 9 being the highest possible costs or rewards that agents can have. We next set the cooperation parameter ϕ to take on integers between -25 and 25, which allows the model to consider extreme cooperative and adversarial cases (see Oey et al., 2022, for related work on adversarial mental-state reasoning). Our model code is available online at `https://osf.io/57n4g`.

5.3 Model Analysis

Our first goal is to use our computational model to test whether our proposal can capture the emergence and use of low-cost communicative blockers (LCCBs). If our model failed to replicate this phenomena, this would imply that our account is incorrect. Specifically, our analyses consist of a set of simulations that test whether the enforcer and decider in our model can reproduce our target phenomena—creating and understanding LCCBs (inspired by those in Figure 5.1).

To explore our model dynamics, we focused on the same simple domain with two doors—door A and door B—and an enforcer that wants the decider to choose door A (Figure 5.2). We assume that the initial scene s_0 has a set of objects between the two doors, such that the objects do not initially block either door, and placing an object in front of either door is equally costly (Figure 5.2a). For simplicity, we also assume that the cost the enforcer incurs in placing an object in front of a door is the same as the cost that the decider incurs when moving that object out of the way. To analyze the core dynamics of the model, we simulated a situation where the enforcer was maximally motivated to affect the decider's behavior (setting their reward to 9), and where the decider was also very cooperative (setting $\phi = 10$). We further set the softmax parameter to a minimum in order to remove any noise in the inferences (setting $\tau = 0.1$). We then tested our model's performance by varying the decider's relative preference for different options.

Figure 5.2a shows a visual depiction of the initial state, and Figures 5.2d-h shows five possible scene transformations that the enforcer could produce (stacking one, three, or four objects in front of door B alone, or also stacking any number of objects

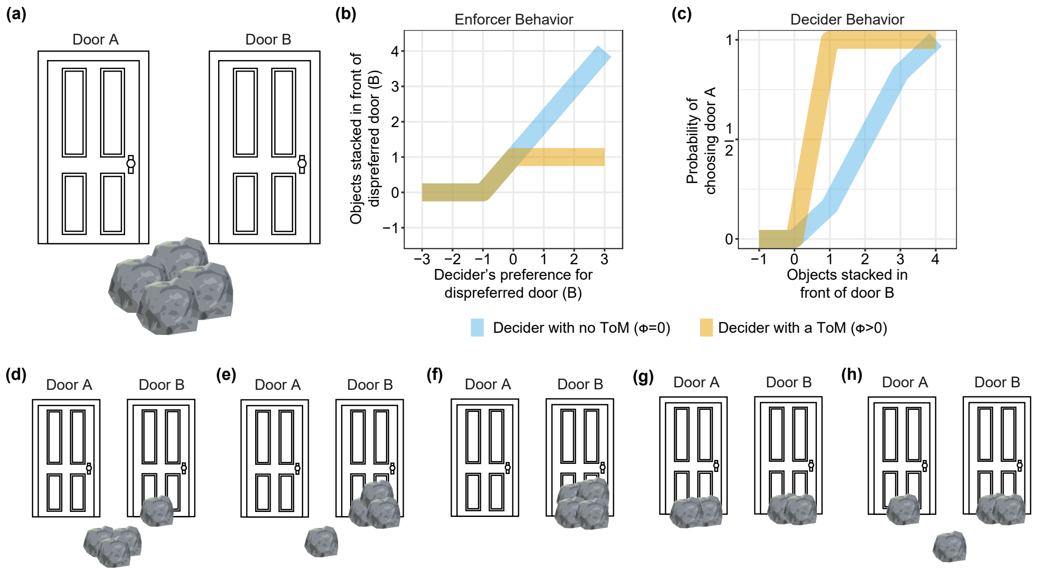

Figure 5.2: Example event used to illustrate model performance. The environment consists of two doors and a stack of boulders between them. The enforcer's goal is to reposition the boulders to get the decider to choose door A. (a) Initial scene state. (b) Enforcer behavior. The x-axis shows the decider's preference for door B (negative values indicating a preference for door A) and the y-axis shows the number of boulders the enforcer stacks in front of door B (negative values indicating stacking objects on door A). The simple enforcer (blue line) builds the smallest possible physical barrier that will dissuade the decider. The complex enforcer (yellow line) places a single boulder in front of door B, even when the decider has a strong preference for going through it. (c) Deciders with a preference for door B reacting to boulders placed in front of that door. The non-mentalistic decider (blue line) slowly becomes more likely to choose door A as a function of how many boulders are blocking door B. The mentalistic decider (yellow line) recognizes the meaning of a single boulder and adjusts their behavior, immediately forgoing their preferred door B and choosing door A instead. (d-h) Visualization of some of the different scenes the enforcer could produce.

in front of door A). To understand our model behavior, we began by contrasting the simple and complex enforcers. Figure 5.2b shows how many objects each enforcer chooses to stack in front of door B (the door they hope the decider will avoid) as a function of the decider's preference for this door. When the decider already prefers door A (negative decider preference along the x-axis in Figure 5.2b), neither enforcer moves any objects. This reflects the enforcers' confidence that the decider will take door A, making any involvement unnecessary.

When the decider prefers door B (positive decider preference along the x-axis in Figure 5.2b), the enforcers begin to place objects in front of the door, producing two different types of behavior. The simple enforcer expects the decider to choose a door based only on their egocentric costs (how difficult is it to walk through each door?) and rewards (how much does the decider want to walk through each door?). Consequently, this enforcer stacks the minimum number of objects necessary to push the decider's choice towards door A. This is captured in Figure 5.2b, where the blue line shows how the simple enforcer stacks more objects as the decider's preference becomes stronger. This behavior reflects a non-communicative barrier-building strategy, where the enforcer is attempting to make it just hard enough for the decider to cross through door B, with the hope that this added cost will shift their preference towards door A.

In contrast to the barrier-building strategy from the simple enforcer, the complex enforcer places a single object in front of door B (as in Figure 5.2d), even when the decider really prefers that door (yellow line in Figure 5.2b). We interpret this as the kind of communicative strategy that we aim to explain (reminiscent of Figure 5.1): The strategy succeeds not because it imposes a high cost on deciders, but because it efficiently reveals the enforcer's mental states. In these cases, the enforcer knows that the decider's egocentric utilities will favor door B, because the single object imposes a negligible cost. The enforcer nonetheless chooses to place a single object in front of door B because they believe that the decider will infer that they are supposed to take door A instead.

Returning to our motivating examples (Figure 5.1), this behavior resembles actions like placing a plastic chair to mark that a parking spot is taken. Here, a plastic chair does little to prevent someone from using the parking spot: moving the chair out of the way is easy, and the cost is probably insufficient to overcome a driver's desire to find a parking spot. However, the object is effective because it reveals that whoever placed the chair is requesting that their parking spot be respected.

Figure 5.2c shows the behavior of our decider model. The non-mentalistic decider responds to the physical costs alone, becoming more likely to abandon their preferred door as a function of how many objects are blocking it. This is visualized by the blue line in Figure 5.2c, which shows a continuous preference change as a function of the number of objects blocking their preferred door. By contrast, the mentalistic decider shows a sharp discontinuity: A single object in front of their preferred door is enough for them to understand that they should avoid that door. This is visualized by the yellow line in Figure 5.2c, where the decider shows a rapid change in strategy as soon as a single object is in front of their preferred door. Together, these results show how our model gives rise to enforcers who use objects in a communicative manner and deciders who can infer the communicative meaning of these objects.

5.4 Experiment 1: Quantitative model evaluation

Having established that our account can replicate the qualitative use and recognition of low-cost communicative blockers (LCCBs), we next test whether our model's exact inferences match human intuitions. That is, our model predicts quantitative patterns about how strong people's intuitions should be in different displays. If participants can interpret LCCBs, but do so in a different way than our model does, the resulting large discrepancies between our model inferences and participant judgments would falsify our account.

In Experiment 1, participants saw a two-dimensional gridworld of a fruit farm with an entrance, pomegranate groves, pear groves, and a set of boulders placed by a farmer to protect their pomegranates from nearby hikers (farmers corresponding to enforcers and hikers to deciders from our Computational Framework).

We tested participants in two conditions (Figure 5.3). In the *non-mentalistic* condition, hikers believe that the boulders are natural constraints, devoid of social meaning, and farmers plan for how many boulders to place accordingly. We therefore expect participants to infer that, the more boulders the farmer places, the more she believes that hikers want to take the pomegranates. We model this condition using the non-mentalistic decider model (Section 5.2.1) and the simple enforcer model (Section 5.2.2).

In our second condition, the *mentalistic* condition, hikers will always know that the boulders were placed by a farmer, and use the costs imposed by these objects to infer the farmer's preferences. In this condition, a single boulder does not necessarily imply an expectation that hikers do not like pomegranates (as would be implied in the *non-mentalistic* condition). Instead, a single boulder might reveal that the farmer expects hikers to infer that they should stay away and act accordingly. By contrast, if the farmer placed multiple boulders, this would reveal that she expects hikers to prefer pomegranates and be uncooperative (otherwise, a single communicative boulder would have sufficed). We model this condition using the mentalistic decider model (Section 5.2.3) and the complex enforcer model (Section 5.2.4).

All studies were approved by Yale's IRB (protocols "Culture and Cognition" #2000022403 and "Online reasoning" #2000020357). Data collection was obtained in the following experiment order: 3a (meaning inference), 3b, 2, 4, 1, 1 replication, 3a (unusualness ratings), and 3c. Our experimental procedure, stimuli, data, analyses, pre-registrations (for Experiments 1 replication, 2, 3a, and 3c), and supplemental materials are available at `https://osf.io/57n4g`. This manuscript includes all experiments, manipulations, and measures in this line of research.

5.4.1 Participants

80 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk ($n = 40$ per condition; $M = 34.81$ years, $SD = 10.31$ years).

5.4.2 Stimuli

Stimuli consisted of 27 10-by-10 gridworlds, with two fruit groves (pears and pomegranates), a hiker, and a set of boulders (see Figures 5.5a-c for examples). The stimuli were designed by parametrically varying two factors: the distance between the hiker and the groves (i.e., the natural cost of the environment; 5, 7, and 9 squares away) and the number of boulders blocking the pomegranates (i.e., the artificial cost introduced by the farmer; 1, 2, or 3 boulders). The hiker's starting position was randomly selected to be at one of the four corners, and the fruit groves were randomly placed on the two adjacent corners relative to the hiker.

5.4.3 Procedure

Participants read a brief cover story explaining that they would see hikers in different farmlands with pear and pomegranate groves (see Figure 5.3 for paradigm schematic). The farmers, who were absent, did not mind hikers taking pears but they wanted to protect their pomegranates. To achieve this, farmers placed boulders in front of their pomegranate groves (see Figures 5.5a-c for examples). Participants then completed a multiple-choice five-question quiz (see online OSF repository for questions) to ensure they understood the task. Participants that answered at least one question wrong were sent to the beginning of the cover story to try again. Participants that failed the questionnaire twice were not permitted to participate in the study.

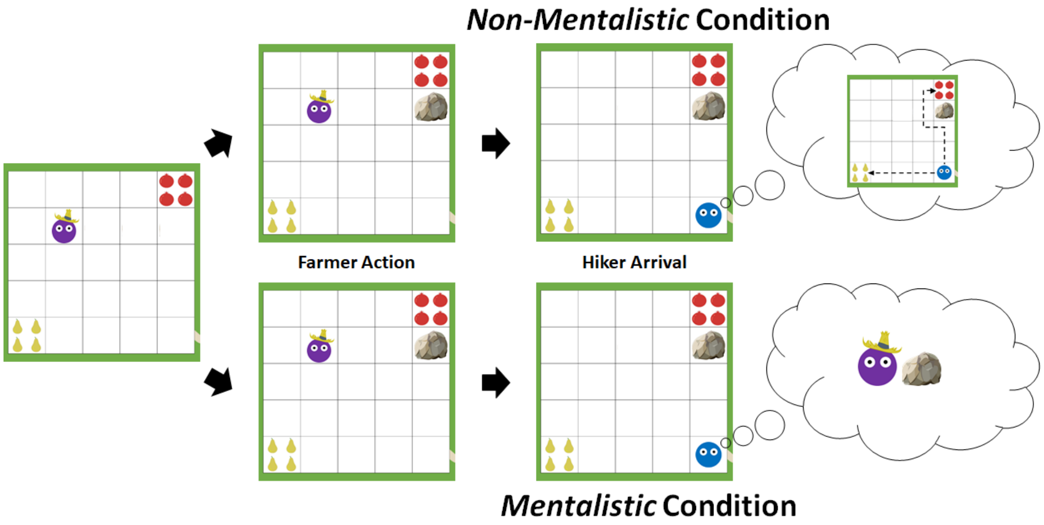

Figure 5.3: Visual schematic of Experiment 1 cover story. Participants learned that a farmer (purple agent) wanted to protect their pomegranates and placed boulders to block the way before leaving. After leaving, a hiker would arrive and decide which fruit to take. In the *non-mentalistic* condition, the hikers treat the boulders as natural constraints, and they therefore decide what to do without thinking about the farmer. In the *mentalistic* condition, the hikers know that a farmer must have placed the boulders, and use this to infer what to do.

Participants in the *non-mentalistic* condition were told that hikers thought the boulders were natural constraints, and that farmers planned how many boulders to place accordingly. That is, the farmer expected hikers to realize that the boulders make it harder to reach a fruit grove, but assume that this was simply a feature of the terrain, rather than an intentional design. In each trial, participants saw an arrangement of boulders and they were asked how much the farmer expected hikers to

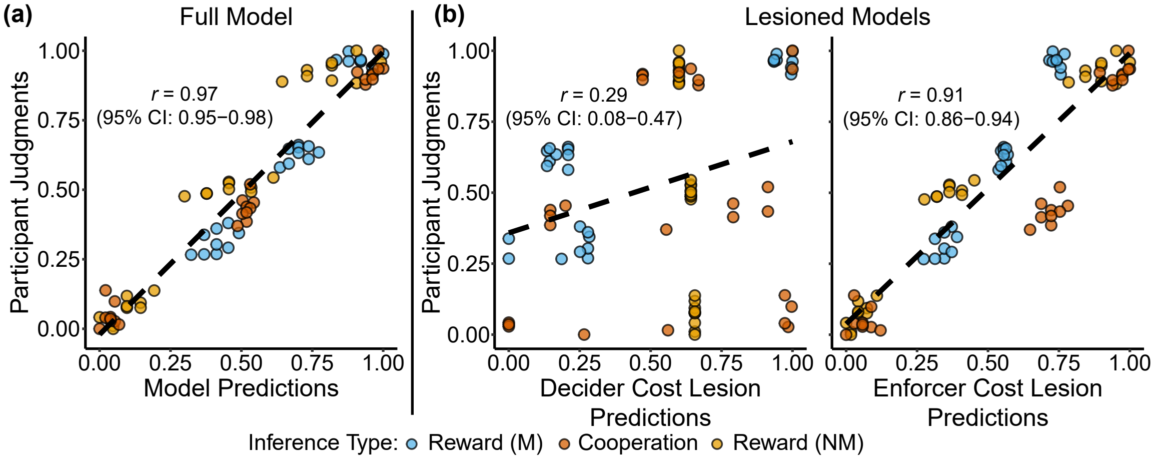

Figure 5.4: Experiment 1 results. Each point represents a judgment, with model predictions on the x-axis and participant judgments on the y-axis. Participants in the *non-mentalistic* condition (NM) condition inferred the hiker's expected preferences, and participants in the *mentalistic* condition (M) additionally inferred the hiker's expected cooperativeness (ϕ). (a) Correlation between our full model and participant judgments. (b) Correlation between lesioned models and participant judgments. Model lesions include removing the influence of cost from the decider (left) and removing the influence of cost from the enforcer (right).

like pomegranates ("How much does this farmer think that hikers like pomegranates?", using continuous sliders ranging from "not at all" to "very much").

Participants in the *mentalistic* condition were told that hikers would always know that a farmer placed the boulders intentionally, and that farmers planned how many boulders to place accordingly. That is, the farmer expected hikers to know that the boulders make it harder to reach a fruit grove, and that these boulders were placed intentionally by someone. In each trial, participants saw an arrangement of boulders and they were asked how much the farmer expected the hiker to like pomegranates. In addition, because the complex enforcer and mentalistic decider include a cooperation parameter ϕ (i.e., the adopted utility weight; Powell, 2022), participants were also asked whether the farmer expected the hiker to be cooperative ("How cooperative does this farmer think hikers are?", using a continuous slider ranging from "not at all" to "very much").

All participants completed the same 27 trials (trial order randomized across participants), where we varied both the initial cost of obtaining each type of fruit (by manipulating the initial distance from the hiker) and the number of boulders that the farmer added (ranging from 1 to 3; see Stimuli).

5.4.4 Model Predictions

Our model's parameters were set prior to data collection (and reflected in the preregistration of the Experiment 1 replication; see Section 5.2.5 and SM at `https:`

//osf.io/57n4g/ for details). For each dependent variable in our task we computed our model's posterior predictive distribution, and used the expected value as the final model prediction.

5.4.5 Results

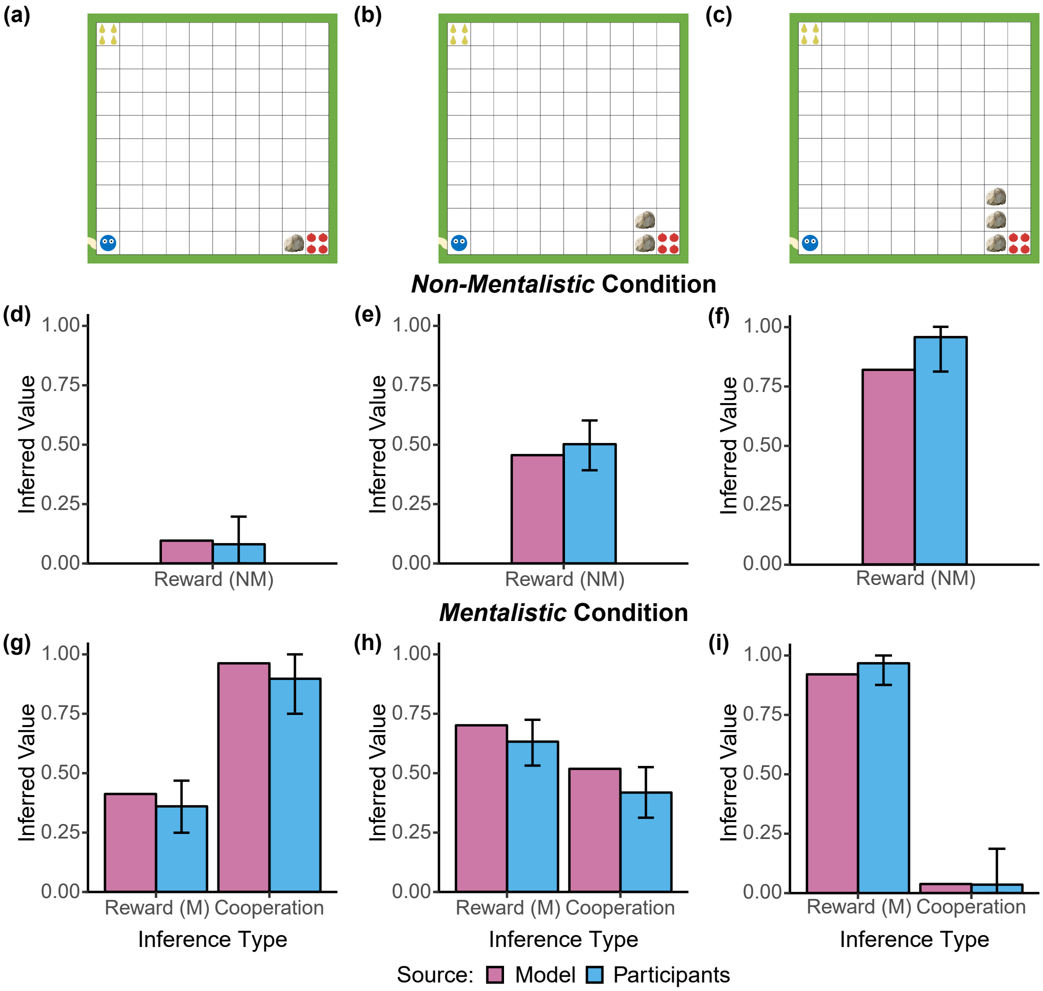

Figure 5.5: (a-c) Example stimuli from Experiment 1. In these examples, both fruit groves were equally far and only varied on the number of boulders a farmer placed. (d-f) Model predictions and participant judgments from the *non-mentalistic* condition in purple and blue, respectively. (g-i) Model predictions and participant judgments from the *mentalistic* condition in purple and blue, respectively. Inference type is along the x-axis and the inferred value is along the y-axis. Error bars are bootstrapped 95% CIs.

Our model and participant judgments showed an overall correlation of $r = 0.97$ ($\text{CI}_{95\%}$: $0.95 - 0.98$; Figure 5.4a). A pre-registered replication of this study produced identical results ($r = 0.98$; $\text{CI}_{95\%}$: $0.96 - 0.98$; see SM for details). The fact that our model captures the fine-grained structure of people's inferences suggests that their inferences resembled the ones obtained by reasoning about the farmer's desires via recursive social reasoning.

Figure 5.5 shows three example trials that highlight the inferences that our model and participants made. In Figures 5.5a-c, the hiker's distance to pomegranates and pears is matched (making the initial cost identical) and the number of boulders in front of the pomegranates varies from 1 to 3. In the *non-mentalistic* condition, the number of boulders should reveal how much the farmer thinks hikers will want to get the pomegranates (because the purpose of the boulders is to introduce a physical cost that outweighs the hikers' desires). Figures 5.5d-f show this effect in both our model and participant reward inferences (with each plot corresponding to the stimuli directly above it; e.g., panel (d) corresponding to map (a)). As the number of boulders increased, participants and our model inferred a stronger preference for pomegranates.

In the *mentalistic* condition, the boulders not only impose a physical cost, but allow hikers to infer the farmer's preferences. Therefore, a single boulder (Figure 5.5g) does not necessarily imply that hikers must not like pomegranates that much (as it did in the *non-mentalistic* condition; Figure 5.5d). Instead, the farmer may have used a single boulder to reveal that they did not want hikers to take the pomegranates. Consistent with this, both participants and our model inferred that hikers could have a higher desire for pomegranates (compare Figures 5.5g and 5.5d), but were highly cooperative. That is, participants and our model inferred that a single boulder was effective because it revealed the farmer's preferences to cooperative hikers (despite its cost not being high enough to outweigh their preferences). This reward difference across conditions is further visualized in Figure 5.6, and was significantly different across conditions ($\Delta R = 0.24$; $p < 0.001$ from a two-tailed t-test).

When the number of boulders blocking the pomegranates increases (Figures 5.5h-i), the farmer's additional actions (placing more than one boulder) can be explained by inferring that hikers must really want the pomegranates and not be particularly cooperative (given that they will be able to infer that the farmer wants them to stay away). Consistent with this, both our model and participants infer a stronger hiker desire and a lower cooperativeness as the number of boulders increases (see SM for additional results of a linear mixed-effects regression predicting these participant reward inferences as a function of boulder count and condition).

Alternative models

While our model captured participant inferences with quantitative accuracy, it is possible that participants reached similar inferences through simpler mechanisms. To test this, we considered two alternative models. A first possibility is that people focus only on an object's position, without considering the costs that it might impose on observers. In our experiment, this means that hikers do not consider the cost of navigating around boulders. We call this model the *Decider Cost Lesion* as it is similar to our model with the difference that it does not reason about the cost that objects impose on deciders. Figure 5.4b (left) shows how this model was no longer able

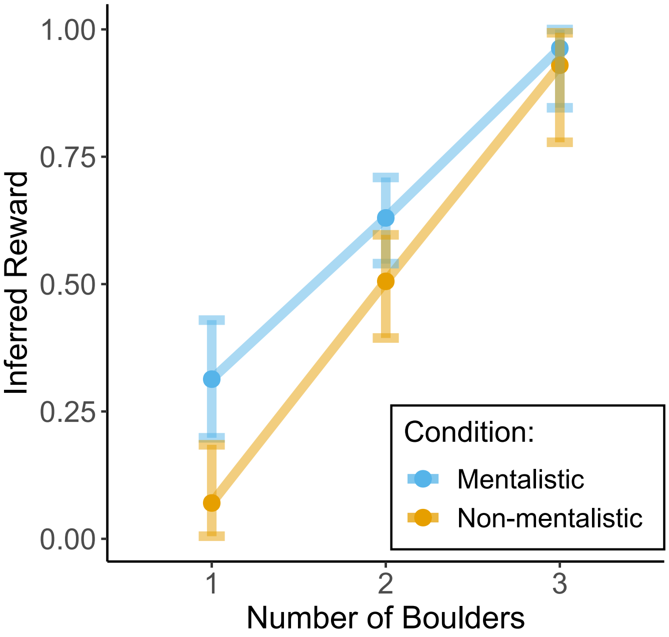

Figure 5.6: Reward inferences across the *non-mentalistic* condition and *mentalistic* condition in Experiment 1. The number of boulders placed by the farmer is on the x-axis and the reward participants inferred is on the y-axis. Error bars are bootstrapped 95% CIs.

to explain participant judgments ($r = 0.29$; $\mathrm{CI}_{95\%}$: $0.08 - 0.47$), and was also reliably worse than our main model ($\Delta r = 0.68$; $\mathrm{CI}_{95\%}$: $0.49 - 0.89$). This result confirms that the cost imposed on deciders is critical for capturing human-like inferences.

A second possibility is that people do consider the costs that an object imposes on their actions (e.g., detecting that an object is making it harder for them to get a certain fruit), but they do not consider the effort that someone had to incur in positioning the object. In our experiment, this means that people do not think about the cost farmers incur when placing boulders. We call this model the *Enforcer Cost Lesion* as it is similar to our model with the difference that it does not reason about the cost the enforcer incurs. Figure 5.4b (right) shows how this lesioned model compares to participant judgments. Although this model performed worse than our main model ($\Delta r = 0.06$; $\mathrm{CI}_{95\%}$: $0.03 - 0.10$), it was nonetheless able to capture the pattern of inferences about deciders quite well ($r = 0.91$; $\mathrm{CI}_{95\%}$: $0.86 - 0.94$).

These results suggest that participant inferences may not depend as heavily on the cost incurred by the enforcer (i.e., the farmer). We believe this is intuitive for the situations that we focused on. For instance, when encountering a broom positioned directly across a door, we intuitively focus on the cost that the broom imposes on us, rather than thinking about the cost the enforcer incurred. At the same time, Figure 5.4b (right) reveals that this lesioned model nonetheless fails to capture a subset of participant intuitions that our main model was able to capture. Specifically, this lesioned model over-estimated hikers' cooperativeness when compared to humans in trials with two boulders (visualized as a cluster of orange points that falls most distant from the best-fit line in Figure 5.4b, right). This is because, according to this lesioned model, placing three boulders is as easy as placing two boulders. Therefore, the farmer choosing not to place an extra boulder at no cost would only be reasonable if the hikers were highly cooperative, to the point that placing a single boulder was guaranteed to be as effective as placing more boulders to block the way. Together, this analysis suggests that the cost incurred by enforcers is less critical for capturing how we infer the communicative meaning of an object, but that people are nonetheless sensitive to it, and use it to infer other agents' cooperativeness. Overall, these lesions show how considering the cost that enforcers incur in positioning objects, and how these objects also impose a cost on deciders, are key to explaining how participants reasoned about objects in our experiment.

5.5 Overview of Experiments 2-4

Our model analyses and Experiment 1 show that people can derive inferences that are quantitatively similar to those from our model. While these results show that people can make these types of social inferences, they do not imply that this is what people do when encountering communicative objects. In Experiments 2-4, our goal is therefore to test an alternative hypothesis: Could simple conventions without inference explain the use of communicative objects in their entirety? That is, conventional knowledge is undoubtedly critical to the everyday use of communicative objects. Our goal is therefore not to question its importance, but to ask what happens when conventional knowledge is unavailable, such as when we encounter an unfamiliar object that might have a communicative purpose. In these cases, do people rely on social inferences like the ones we proposed? Or are they unable to make any conclusions given the absence of an explicit convention?

Given that people will use conventional knowledge when it is available, our experiments here focus on objects that are not associated with a pre-existing communicative meaning. We begin by testing two predictions. First, if the meaning of communicative objects were based on explicit convention alone, then people should detect an object as communicative only when they have been explicitly taught about its meaning (therefore falsifying our account). By contrast, our account predicts that people should be more likely to associate low-cost novel objects with a communicative purpose, relative to novel objects that impose no cost (as these fail to reveal the mental states of whoever positioned the object). We test this prediction in Experiment 2. Second, if the meaning of objects were driven by explicit pedagogy and convention alone, then people should be unable to infer the meaning of a novel object, even when they know that the object has a communicative purpose (therefore falsifying our account). By contrast, our account predicts that people should be able to infer the communicative meaning of an object when its placement (i.e., the cost it imposes) reveals the enforcer's mental states. We test this prediction in Experiment 3. Finally, if people are engaging in social inferences to infer the meaning of novel objects, our work brings forth the question of how quickly these meanings might become conventionalized. In Experiment 4, we test the idea that people might be able to quickly treat the meaning of a novel object as conventional.

5.6 Experiment 2: Are objects that impose a cost more likely to be communicative?

If people's reasoning about low-cost communicative blockers is driven entirely by explicit object-meaning conventions, then people should report that an object is communicative only when they have been explicitly taught its meaning. In Experiment 2, we therefore tested whether people believe that objects that impose a low cost on deciders are more likely to be communicative relative to objects that do not impose a cost, as our account predicts but the explicit convention account does not.

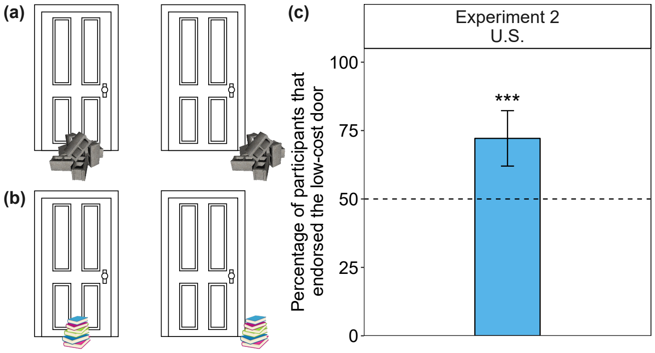

Figure 5.7: (a-b) Example stimuli from Experiment 2. In both panels, the left side shows a low-cost door and right side shows a no-cost door. (c) Experiment 2 results. The blue bar represents the percentage of participants that associated the low-cost door with having a communicative purpose. Error bars are bootstrapped 95% CIs.

5.6.1 Participants

80 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk (M = 37.84 years, SD = 12.22 years). 14 additional participants were recruited and replaced for failing our inclusion criteria (see Results).

5.6.2 Stimuli

Stimuli consisted of eight images of pairs of doors, with each pair consisting of a "low-cost door" and a "no-cost door" (e.g., Figures 5.7a-b). Each of the eight pairs was associated with one of eight objects that are not conventionally used to communicate: a plant, a chair, a pile of books, a pile of cinderblocks, some tape, some meter sticks, a hat, and a fishbowl tied to a tack on a door frame (see online OSF repository for the full stimuli set). In the low-cost doors, the object was placed directly in the middle of the doorway (e.g., Figures 5.7a-b, left), and in the no-cost doors, the object was placed next to the door, not blocking the way (e.g., Figures 5.7a-b, right). Half of the door pairs were open and the other half were closed.

5.6.3 Procedure

Participants were asked to imagine leaving an office and encountering a pair of doors, each with an object nearby. Participants then answered a simple multiple-choice attention-check question ("What objects are in front of the doors?"). Participants that answered incorrectly were sent to the beginning of the cover story and not permitted to access the experiment until they answered correctly.

Participants then saw a single trial containing a low-cost door and a no-cost door, both with the same object nearby (see Figures 5.7a-b for examples; door order randomized across participants). Participants were asked, "Which door do you think someone was trying to tell you something?", followed by a manipulation check ("Which door requires more work to walk through?") and an inclusion question ("Do you think you would be able to walk through this door if you wanted to?"). These questions were always presented in the same order (see online OSF repository for the full procedure details).

5.6.4 Results

Participants who did not think they could walk through the doors were excluded from the study and replaced (as our interest is in the inferences people make when objects are not seen as insurmountable physical constraints; $n = 14$; 14.89% exclusion rate). Of our final sample, 72.50% of participants reported that the low-cost door was more likely to be communicative ($\text{CI}_{95\%}$: $62.50\% - 82.50\%$; $p < 0.001$ from a two-tailed binomial test; Figure 5.7c), rather than performing at chance, as expected by the explicit pedagogy account (see SM for a supplemental analysis confirming this result). Our exclusion rate (14.89%) was lower than recent estimates of attentiveness on Mechanical Turk (estimated to be at approximately 20%; Arechar and Rand, 2021), suggesting that these participants were simply inattentive. However, these participants showed the same qualitative pattern of responses as those included in the task (see SM for details on excluded participants). It is therefore possible that these participants were attentive but did not interpret our inclusion question ("Do you think you would be able to walk through this door if you wanted to?") as referring to physical plausibility alone, integrating social expectations as well (given the heavy social focus of the task).

5.7 Experiment 3a: Are low-cost objects interpreted as communicative blockers?

Having found that people are more likely to interpret a low-cost object as communicative, in Experiment 3a, we next test what meaning people are more likely to associate with it. If the object is unfamiliar, the explicit convention account predicts that participants should be at a loss about what it means, given the absence of an established object-meaning mapping. Instead, if people are making inferences about why someone would place the object to impose a cost, they should infer that the object is more likely to mean that they should avoid the door.

5.7.1 Participants

160 U.S. participants (as determined by their IP address) were recruited via Amazon Mechanical Turk (n = 80 per condition; M = 34.85 years, SD = 8.38 years). 17 additional participants were recruited and replaced for failing our inclusion criteria (see Results).

5.7.2 Stimuli

Stimuli consisted of 16 images of pairs of doors, with each pair consisting of an empty door and a door with an object nearby (e.g., Figures 5.8a-b; using the same objects from Experiment 2). In the "low-cost pair", the object was placed directly in the middle of one of the doorways (e.g., Figure 5.8a), and in the "no-cost pair", the object was placed next to one of the doors, not blocking the way (e.g., Figure 5.8b). Half of these door pairs were open and the other half were closed.

5.7.3 Procedure

The procedure was similar to Experiment 2. Participants were asked to imagine leaving an office building and finding two identical exits. Participants then saw that one of the doors had an object nearby, and they were told that it was unclear whether someone wanted them to take that door or to avoid it. Participants were then asked a multiple-choice attention-check question: "What is the only difference between the two exits?" Participants next saw a pair of doors (either a low-cost pair for participants randomly assigned to the *low-cost* condition or a no-cost pair for participants randomly assigned to the *no-cost* condition; door order randomized across participants), and were asked: "What do you think someone was trying to tell you about the door with the OBJECT?" (possible responses: "You should walk through the door with the OBJECT" or "You should not walk through the door with the OBJECT"). Participants then responded to the same manipulation-check and inclusion questions from Experiment 2.

5.7.4 Results

Participants who said the empty door was harder to walk through were excluded from the study and replaced ($n = 17$; 9.60% exclusion rate). Of our final sample, 87.50% of participants in the *low-cost* condition inferred that they were supposed to avoid the door with the object ($\text{CI}_{95\%}$: $80.00\% - 93.75\%$; $p < 0.001$ from a two-tailed binomial test). By contrast, only 60.00% of participants in the *no-cost* condition inferred that this door should be avoided ($\text{CI}_{95\%}$: $48.75\% - 70.00\%$; $p = 0.093$ from a two-tailed binomial test[1]), a proportion not significantly different from chance. Moreover, the number of participants inferring that they should avoid the door was significantly higher in the *low-cost* condition relative to the *no-cost* condition ($p < 0.001$ by Fisher's exact test). See SM for additional results showing that this effect cannot be explained by appealing to the idea that people believe that more 'unusual' arrangements of objects are more likely to signal avoidance. The fact that participants did not perform at chance in both conditions suggests that people do not rely purely on conventional knowledge. The pattern of results from our excluded participants was qualitatively consistent with that of participants included in the task (see SM for details on excluded participants).

5.8 Experiment 3b: Replication with the Tsimane'

Experiment 3a suggests that people's reasoning about low-cost communicative blockers cannot be reduced to explicit object-meaning conventions. However, it is possible that these inferences are culture-specific. Because our model is built on simple aspects of human cognition that are thought to be universal, the absence of these inferences in other cultures would challenge our account. As a first step in exploring this possibility, we replicated a variation of Experiment 3a with the Tsimane'—a farming-foraging group native to the Bolivian Amazon. The Tsimane' live in non-industrialized communities along the Maniqui river and have less exposure to market-integrated communities compared to U.S. participants. Comparing the Tsimane' and WEIRD participants (Western, educated, and from industrialized, rich, and democratic countries; Henrich et al., 2010) has helped identify cultural influences in color-word vocabulary (Conway et al., 2020; Gibson et al., 2017) and music perception (McDermott et al., 2016), and has also helped rule out cultural influences in other domains, such as the stages of number-word learning in children (Piantadosi et al., 2014; Jara-Ettinger et al., 2017) and the ways in which people identify communicative action (Royka et al., 2022). We therefore sought to test the Tsimane' as a way to explore if these inferences also emerge in a culture that is substantially different from the U.S.

[1] Our pre-registered analysis proposed to use logistic regressions to study this effect. We instead present binomial tests for clarity, but the results are identical under our pre-registered analysis and can be found in SM.

5.8.1 Participants

133 Tsimane' adults were recruited in their local communities in the Bolivian Amazon (M = 33.12 years, SD = 15.40 years). 17 additional participants were recruited but excluded from the study for failing to complete the study (see Results).

5.8.2 Stimuli

Stimuli consisted of six images of doors, each with an object in front of it (e.g., Figures 5.7a-b). We used a subset of the objects used in Experiment 3a that Tsimane' participants were familiar with (as determined by our interpreters) while remaining unconventional as communicative objects: a plant, a chair, and a pile of cinderblocks. Each object was associated with two different doors: a "low-cost door", where the object was placed directly in the middle of the doorway (e.g., Figure 5.7a, left), and a "no-cost door", where the object was placed next to the door, not blocking the way (e.g., Figure 5.7a, right). Half of these doors were open and the other half were closed.

5.8.3 Procedure

The procedure was adapted from Experiment 3a to be more intuitive for our participants, based on feedback from our interpreters. Participants were asked to imagine deciding to enter a friend's house through one of two possible doors. Participants were then shown a low-cost door and a no-cost door sequentially (order counterbalanced across participants, with different objects used for each door) and were asked: "Do you think the owner wants you to enter or stay away?"

5.8.4 Results

Participants that failed to complete both trials were excluded from the study (n = 17; 11.33% exclusion rate; see SM for details on excluded participants). Like U.S. participants, Tsimane' participants inferred that they should avoid a door when the object was minimally blocking the door (85.71%; $\mathrm{CI}_{95\%}$: $79.70\% - 91.73\%$; $p < 0.001$ from a two-tailed binomial test), but not when the object was on the side of the door (30.08%; $\mathrm{CI}_{95\%}$: $22.56\% - 38.35\%$; $p < 0.001$ from a two-tailed binomial test).

5.8.5 Discussion

These results suggest that, like U.S. participants, Tsimane' participants also inferred avoidance from objects that impose a low cost. Critically, this experiment used objects that were familiar to the Tsimane', but not typically used by them to communicate (as determined by our local interpreters). This approach followed the same logic as our design with U.S. participants, which also used familiar objects but, critically, not ones typically used to communicate. This enabled us to maximize *equivalence* (Matsumoto and Yoo, 2006; Van de Vijver and Leung, 2021; Poortinga, 1989)—the goal of reaching

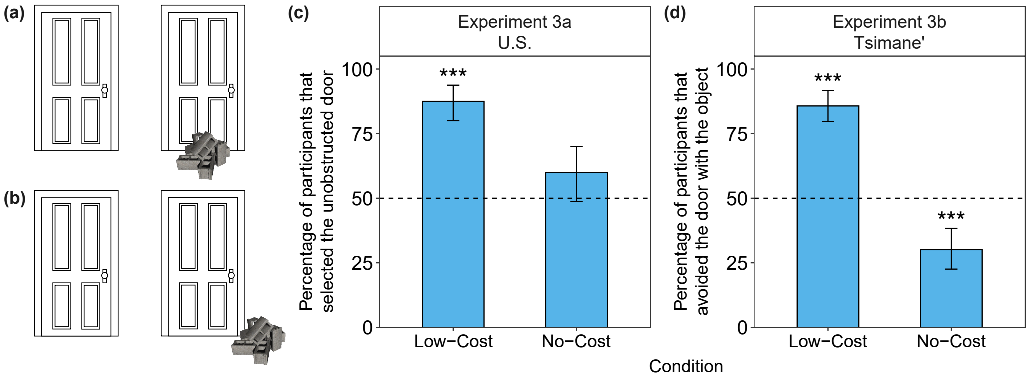

Figure 5.8: (a) Example stimuli from the *low-cost* condition in Experiment 3a. (b) Example stimuli from the *no-cost* condition in Experiment 3a. (c) Experiment 3a results. The blue bars represent the percentage of U.S. participants that selected the empty door as a function of condition. (d) Experiment 3b results. The blue bars represent the percentage of Tsimane' participants that inferred that they should not go through the door as a function of door type. Error bars are bootstrapped 95% CIs.

similarity in conceptual meaning across groups to support meaningful comparisons. At the same time, a stronger test of our hypothesis would have included entirely novel objects, which would have allowed us to test the nature of these inferences without any possible influence from prior object knowledge. These results, therefore, only provide evidence that people can infer the communicative meaning of familiar objects that are not typically communicative, and we do not know if these inferences would extend to entirely novel objects.

5.9 Experiment 3c: Inferences from conventional communicative blockers

In Experiments 3a and 3b, we found evidence that people are sensitive to the cost an object imposes when reasoning about its potential communicative meaning. Importantly, these experiments used objects with no pre-existing communicative meaning associated with them. Under our account, these inferences become critical when people do not have a pre-existing convention, but may become less important when they already know an object's communicative meaning. In Experiment 3c, we replicated Experiment 3a using conventional objects. If people are constantly making cost-based inferences with all communicative objects, these results should replicate the pattern of Experiment 3a: inferring avoidance in the *low-cost* condition, but not in the *no-cost* condition. However, if these inferences are only at work when encountering novel objects, people should report the conventional communicative meaning of the object regardless of the cost that it imposes.

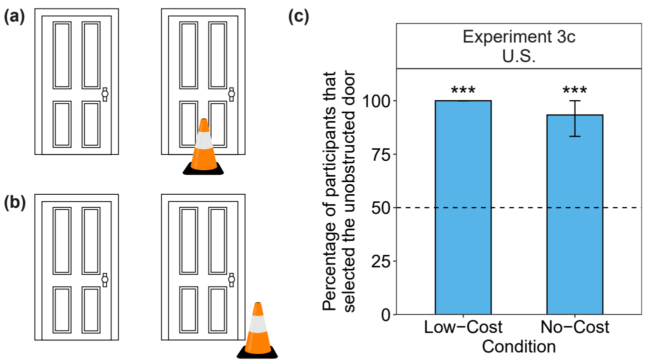

Figure 5.9: (a) Example stimuli from the *low-cost* condition in Experiment 3c. (b) Example stimuli from the *no-cost* condition in Experiment 3c. (c) Experiment 3c results. The blue bars represent the percentage of U.S. participants that selected the empty door as a function of condition. Error bars are bootstrapped 95% CIs (not visible in *low-cost* condition due to participants performing at ceiling).

5.9.1 Participants

60 U.S. participants (as determined by their IP address) were recruited via Prolific ($n = 80$ per condition; $M = 35.72$ years, $SD = 12.48$ years). 4 additional participants were recruited and replaced for failing our inclusion criteria (see Results).

5.9.2 Stimuli

Stimuli consisted of six images of pairs of doors, with each pair consisting of an empty door and a door with an object nearby (similar to those used in Experiment 3a, but using different objects; e.g., Figures 5.9a-b). Here we used objects that are conventionally used as communicative blockers: a traffic cone, construction tape, and a stanchion.

5.9.3 Procedure

The procedure was identical to that used in Experiment 3a, with the only difference being the objects that participants saw. Participants then responded to the same manipulation-check and inclusion questions from Experiment 3a.

5.9.4 Results

Participants who said the empty door was harder to walk through were excluded from the study and replaced ($n = 4$; 6.25% exclusion rate), since the empty door is never harder to walk through (see SM for details on excluded participants). Of our final sample, 100.00% of participants in the *low-cost* condition inferred that they were supposed to avoid the door with an object (CI$_{95\%}$: 100.00% – 100.00%; $p < 0.001$

from a two-tailed binomial test[2]. 93.33% of participants in the *no-cost* condition also inferred that the door should be avoided ($\text{CI}_{95\%}$: $83.33\% - 100.00\%$; $p < 0.001$ from a two-tailed binomial test. While the number of participants inferring that they should avoid the door was qualitatively higher in the *low-cost* condition relative to the *no-cost* condition, this difference was not significant ($p = 0.492$ by Fisher's exact test). These results suggest that conventional knowledge was not driving our effect in Experiment 3a, and that people may have a lower reliance on costs when interpreting conventional communicative objects.

5.10 Experiment 4: Conventionalizing object meanings

Experiments 2, 3a, and 3b suggest that people are sensitive to costs when reasoning about an object with no pre-existing convention about its meaning. Experiment 3c further shows evidence that this sensitivity disappears when the object already has a conventional meaning associated with it. Experiment 4 presents an initial test of how objects that trigger inferences might become conventionalized (no longer requiring inference).

The process of associating objects with meaning might be particularly valuable, as it would help people minimize cognitive demands (Birch and Bloom, 2007; Back and Apperly, 2010; Keysar et al., 2000) and cognitive effort (Shenhav et al., 2017; Kool and Botvinick, 2018). That is, people might do social inference to interpret novel communicative objects, but treat them as conventional quickly afterwards. This hypothesis would help explain why some communicative objects, like chairs to indicate shoveled parking spots (e.g., Figure 5.1c), are used consistently (although not always; e.g., Figure 5.1i), and is parsimonious with a resource-rational view of the mind (Lieder and Griffiths, 2019; Griffiths et al., 2015).

Participants in Experiment 4 were first asked to infer the meaning of a low-cost object, and then completed a second trial that either showed a door with a picture of the same object (*congruent* condition) or a picture of a new object (*incongruent* condition). The critical idea in this experiment is that the picture in the second trial never imposes a cost. If participants interpret that picture independently, they should perform at chance when asked what it means (replicating the responses in the *no-cost* condition of Experiment 3a). Alternatively, people might infer the meaning of the low-cost object in the first trial and then immediately treat it as conventional. If so, then participants should report the same inference when they see a picture of the same object (i.e., in the *congruent* condition) despite the object not imposing a cost, but not when they see a picture of an unrelated object (i.e., in the *incongruent* condition).

[2] Our pre-registered analysis proposed to use t-tests to study this effect. We instead present binomial tests for clarity, but the results are identical under our pre-registered analysis and can be

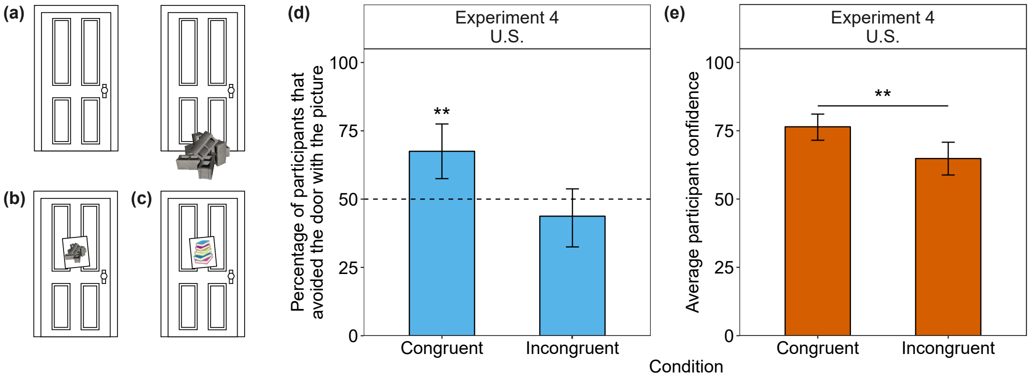

Figure 5.10: (a) Example stimuli from the first trial of Experiment 4. (b-c) Example of a door from the *congruent* and *incongruent* condition relative to (a), respectively. (d) Percentage of participants that inferred that they should avoid the door with the picture as a function of condition in Experiment 4. (e) Average reported confidence rating in this inference. Error bars are 95% bootstrapped CIs.

5.10.1 Participants

160 U.S. participants (as determined by their IP address) were recruited using Amazon Mechanical Turk (n = 80 per condition; M = 36.13 years, SD = 10.90 years). 42 participants were recruited and replaced for failing our inclusion criteria (see Results).

5.10.2 Stimuli

Stimuli consisted of 16 images of pairs of doors, with each object (using the same eight objects from Experiments 2 and 3a) associated with two pairs. In the "low-cost pair", the object was placed directly in the middle of one of the doorways (e.g., Figure 5.10a), next to an empty door. In the "symbol pair", a picture of the object was placed directly in the middle of one of the doorways (e.g., Figures 5.10b-c), next to an empty door. Half of these door pairs were open and the other half were closed (with the picture hanging from the top of the door frame when the door was open).

5.10.3 Procedure

The experiment began in an identical way to Experiment 3a, with the difference that the cover story named the person who had positioned the communicative object. This was done so that we could ensure participants understood that the communicative objects came from the same agent across trials (name randomized across participants).

In the first trial, participants saw a low-cost pair of doors (e.g., Figure 5.10a) and were asked: "What do you think NAME was trying to tell you about the door with the OBJECT?" (with the same possible responses from Experiment 3a) and "How

found in SM.

confident are you that that's what NAME was trying to tell you?" (using a continuous slider ranging from "not confident at all" to "very confident").

Participants were told that their inference was correct (regardless of their answer) and they were next presented with a symbol pair of doors. In the *congruent* condition, one of the doors had a picture of the object from the previous trial (e.g., Figure 5.10b after seeing Figure 5.10a, right, on the first trial). In the *incongruent* condition, one of the doors had a picture of a new object (e.g., Figure 5.10c after seeing Figure 5.10a, right, on the first trial). Participants were asked the same two questions from the first trial, followed by the manipulation-check and inclusion questions from Experiments 2 and 3a.

5.10.4 Results

Participants who did not respond that the object in the first trial imposed a cost were excluded from the study and replaced ($n = 42$; 20.79% exclusion rate), as our interest is in how participants generalize the inferred meaning of objects that they perceived as imposing a cost (see SM for details on excluded participants).

The first trial replicated our results from Experiment 3a, with 66.25% ($\text{CI}_{95\%}$: 58.75% – 73.75%) of participants inferring that they should avoid the door with the object ($p < 0.001$ from a two-tailed binomial test). Participants reported an average confidence rating of 76.96% ($\text{CI}_{95\%}$: 73.71% – 80.03%).

We next turned to our main question of interest. If participants treat the picture in the second trial as a novel signal, they should perform at chance in both conditions, as the picture does not impose a cost. By contrast, if participants assume that the signaler was treating the object as a new convention, they should infer that the door should be avoided in the *congruent* condition, but perform at chance in the *incongruent* condition. As predicted, participants in the *congruent* condition judged that the door with the picture should be avoided, despite the picture not imposing a cost (67.50%; $\text{CI}_{95\%}$: 57.50% – 77.50%; $p < 0.01$ from a two-tailed binomial test). In the *incongruent* condition, only 43.75% of participants inferred that the door should be avoided ($\text{CI}_{95\%}$: 32.50% – 55.0%; $p = 0.314$ from a two-tailed binomial test). Moreover, participants were significantly more confident in their interpretation of the picture in the *congruent* condition (76.45%; $\text{CI}_{95\%}$: 71.46% – 81.09%) relative to the *incongruent* condition (64.83%; $\text{CI}_{95\%}$: 58.78% – 70.81%; $W = 3978$, $p < 0.01$ from a U-test).

5.11 Experiments 2-4 Discussion

Experiments 2-4 suggest that people can infer the potential communicative meaning of an object in the absence of explicit pedagogy and convention. People's inferences were qualitatively consistent with our proposal, where communicative inferences are guided by reasoning about others' mental states. Is it possible that participants arrived to these inferences through simpler heuristics?

A first concern is that an object placed in front of a door might be more salient than an object placed to the side. It is possible that participants in Experiment 2 thought that visually-salient objects (rather than low-cost objects) were more likely to be communicative. However, that it is unclear why people would expect visual salience to imply avoidance, and there are cases where visual salience is interpreted to mean the opposite (Misyak et al., 2016). In addition, a general expectation that communicative objects should be visually salient is not mutually exclusive with our account, and is in fact consistent with the main idea in our proposal: To elicit mental-state inferences through objects, the communicator must also ensure that the recipient will notice the object in the first place.

A second potential concern is that people find unusual arrangements of objects to be more likely to be communicative, and to signal avoidance. To test for this possibility, in Experiment 3a we collected "unusualness" ratings for our stimuli and found that the cost information significantly explained participant judgments when controlling for unusualness (see SM for details).

5.12 General Discussion

Human communication is remarkable in that, beyond words and gestures, we can communicate through objects. Here we proposed that this ability emerges from our capacity to represent and infer other people's mental states. Specifically, we proposed that, if we can reason about other people's mental states based on how they manipulate objects in the environment, then people can also arrange objects with the purpose of revealing their mental states to agents who encounter these objects.

In exploring this idea, our paper makes three contributions. First, we implemented a computational model of social inference from physical objects. This model revealed how communicators can use objects to elicit mental-state inferences in observers, and how observers can infer the communicative meaning of objects, without any direct communication occurring between the agents. This provided proof-of-concept that the computations proposed in our account are sufficient to give rise to this phenomena. Second, we directly tested whether people's inferences about the communicative meaning of an object could be explained by our account. Finally, we tested whether people's intuitions about communicative objects could only be the result of explicit systems of pedagogy and convention. Combined, these results suggest that mental-state inferences support the creation and interpretation of novel communicative objects, while pedagogy and convention drive their widespread use.

5.12.1 Model assumptions and study limitations

At its core, our model consisted of a Theory of Mind implementation which we built following several computational principles that each enjoy strong empirical support:

social interactions involve agents thinking about each other's mental states (Goodman and Frank, 2016; Frank and Goodman, 2012); mental-state reasoning is structured around an assumption that agents maximize utilities (Jara-Ettinger et al., 2016; Jern et al., 2017; Lucas et al., 2014); inferences about other people's minds is performed via some approximation of Bayesian inference (Baker et al., 2009, 2017); and agents are typically cooperative, particularly when cooperation is easy (Powell, 2022; Rand, 2016; Warneken and Tomasello, 2006). To capture how we might extract social information from the physical world, we made three further assumptions in our experiments and model.

Our first assumption was that agents can identify which objects in a scene were manipulated by an agent (as opposed to being the result of some inanimate force; e.g., the wind). This assumption was reflected in the formulation of our decider model, where deciders know both the initial scene s_0 and the final scene s (such that any discrepancy between s and s_0 reveals the costs that the enforcer incurred, and the costs that they introduced to the decider). While related research shows that some form of this capacity emerges in infancy (Newman et al., 2010; Keil and Newman, 2015b), it is likely that in more realistic contexts people do not know the initial scene s_0 and instead have a prior distribution over potential initial scenes $p(s_0)$.

Our second assumption was that the objects we considered were intentionally placed by an agent. In real-world situations, this is not known a priori and people must determine which objects were placed intentionally and which were not. Intuitively, there are many objects that, when encountered in front of a door, would not elicit a communicative inference because their placement appears unintentional. For instance, finding an empty paper bag (which an agent might have simply discarded), a soccer ball (which might have rolled over to the front of the door), or a wallet (which an agent might have dropped) would not trigger communicative inferences, because their placement does not seem intentional. Recent research has found that people see behavior as intentional when the outcome causally depends on the agent's desires (Quillien and German, 2021) and when the agent's behavior increases the odds of the outcome happening (Ericson et al., 2023). Integrating these types of processes into the detection of communicative objects is a key step towards having a more flexible framework that can both infer the meaning of communicative objects, and disregard objects that lack a communicative purpose.

Our third assumption was that people can estimate the cost associated with moving objects (for the enforcer) and navigating around them (for the decider), but our model did not explicitly capture how people determine these costs. Recent work shows that, from childhood, people might estimate the effort and difficulty of manipulating objects through an intuitive theory of physical action (Yildirim et al., 2019; Gweon et al., 2017), and integrating this work into our model may enable us to explain communicative objects in more complex situations. This is a direction we hope to pursue in future work.

Our work also has two important limitations. Our first limitation is that we focused our analysis on a specific class of objects where agents reveal their mental states by imposing minimal costs on observers, which we referred to as low-cost communicative blockers (LCCBs). These LCCBs are widespread and easy to find in our everyday lives (see Figure 5.1 for examples). How might our framework extend to other types of communicative objects in broader contexts? Most directly, our framework can also explain cases where agents decrease a cost to signal invitation (e.g., leaving a box of cupcakes intentionally open to signal that anyone can grab one, or leaving an office door ajar to indicate we can be interrupted). From our model's perspective, these inferences are symmetrical, with the only difference being that the observer would infer that a communicator intentionally lowered the cost (rather than increased it). This is a prediction that we hope to test in future work.

The most general formulation of our proposal is not intrinsically tied to cost manipulation specifically, but rather to environmental manipulations that reveal mental states. Our work is thus consistent with related work showing that people can use no-cost markers to signal flexible meanings. For instance, when given a sticker to mark which of three cups someone should choose, people use the sticker to signal the right cup. By contrast, when one of the three cups must be avoided, people now use the sticker to indicate avoidance (Misyak et al., 2016). This has been hypothesized to reflect a process known as "virtual bargaining" (Misyak et al., 2016, 2014; Misyak and Chater, 2014), where people produce whichever solution they would have reached if they had gotten the opportunity to have an explicit discussion about it.

More broadly, if people communicate through objects by attempting to elicit mental states in observers, then communicators must also pay attention to additional dimensions that we did not explicitly model. Specifically, for observers to infer mental states from an object's position, the object must be noticeable, and appear intentionally placed. Otherwise, the object might be ignored or dismissed. Intuitively, these features might also play an important role in the use of low-cost communicative blockers (Figure 5.1). These additional dimensions are vital for a full computational model of how we infer mental states from objects.

A second key limitation is that our model represents objects in terms of the utilities they provide or impose on agents. This abstraction means that our model does not make any conceptual distinctions between different types of objects as long as they impose the same costs (or provide the same rewards). For instance, our current implementation represents a plastic cup and an expensive water bottle on a table as equivalent (because they impose the same physical cost), even though these two objects would likely elicit different mental-state inferences in observers (in one case, an observer might assume it is trash, while in the other it may be interpreted as a "spot saver"). Similarly, a delivery bag in front of a door may impose a small cost, but knowing what delivery bags are would prevent us from inferring that this is a low-cost communicative blocker. While these examples demonstrate how representing an object's category can strip it of a communicative purpose, there are also cases where representing an object's category strengthens it. For instance, a sisal rope and red velvet rope can both be used to communicate avoidance, but the latter communicates it more strongly (and also carries the implication that there could be negative consequences if the message was ignored, possibly due to knowing that the object was costly and built specifically for this purpose). Intuitively, agents also strategically take this into account when deciding which objects to use to communicate. This is broadly consistent with our account, as it reveals further social reasoning about what observers might find easier to interpret, but is not yet captured by our model.

5.12.2 Open questions

Our work leaves several major questions open. A first open question stems from our focus on adults: What is the developmental trajectory of how people use communicative objects? Related research shows that the capacities necessary for these kinds of inferences emerge early in development. From early childhood, people can infer the presence of a hidden agent based on the structure of the environment (Newman et al., 2010; Keil and Newman, 2015b; Saxe et al., 2005; Ma and Xu, 2013); we can estimate the difficulty associated with fulfilling different tasks (Gweon et al., 2017; Bennett-Pierre et al., 2018; Yildirim et al., 2019); and we can explain behavior in terms of unobservable mental states, like beliefs, desires, and intentions (Wellman, 2014; Gopnik et al., 1997). Most strikingly, recent work has shown that children can also use Theory of Mind to infer the transmission of ideas based on how different agents build similar artifacts (Schachner et al., 2018; Hurwitz et al., 2019). This "intuitive archeology" likely shares a common basis with the inferences in our model, opening the possibility that even children can detect and infer the communicative meaning of objects through Theory of Mind. It may even be the case that the inferences we studied here are in fact an extension of people's "intuitive archeology" and ability to reason about the history of objects.

A second open question is the relative difficulty in creating versus interpreting communicative objects. While our computational model can explain both the creation and interpretation of communicative objects (as revealed in our model simulations), our experimental work focused exclusively on the second component. We did so because, under our account, all people ought to be able to infer the meaning of communicative objects, while the ability to generate them can be limited to a few individuals. We thus do not know the extent to which people can easily create ad-hoc communicative objects in new situations. It is possible that, when creating new communicative objects, people might err on the side of caution and prefer to make the costs higher than would be necessary, to minimize the chance of an agent not realizing the communicative content. As people become more confident that the objects they place are being recognized as communicative, they may subsequently lower the costs. We hope to explore these questions in future work.

A third open question is whether the inferences that we studied here are an extension of pragmatics in language. People's ability to derive non-literal meaning in language is supported by a form of recursive reasoning, captured by the Rational

Speech Act (RSA) framework (Franke and Jäger, 2016; Scontras et al., 2018; Frank and Goodman, 2012; Goodman and Frank, 2016). Our model can be thought of as an RSA model where the medium people use to communicate is objects (instead of utterances) and our costs are physical effort (rather than memory retrieval or utterance length). At the same time, the ability to perform recursive social inference is not unique to language and it is possible that the phenomena we studied here reflect a more general, non-linguistic Theory of Mind. As such, these phenomena might constitute a more primitive form of communication that precedes the ability to do pragmatics in language (i.e., a kind of proto-communication). This is a question that goes beyond the scope of our work.

Finally, our work leaves open the question of how to capture other types of inferences that people make from objects. Intuitively, communicative inferences are only a sliver of the social information that we can read from objects. For instance, objects can also lead us to infer aesthetic goals (e.g., placing an object that we like in a visible location within our house), functional goals (e.g., leaving objects like winter gloves next to our front door for convenience), or even personality traits (e.g., inferring that someone is messy based on the general pattern of objects on their desk; Gosling et al., 2002). It is possible that these inferences could be captured through a richer model of Theory of Mind that can consider a broader set of goals that agents have when interacting with the environment, and this is a question that we hope to pursue in future work.

5.13 Concluding Remarks

Humans have a remarkable capacity to share their mental states through their behavior, language, and even the way they arrange objects in their environment. Our work shows one way in which people can share their mental states through objects. And yet, the types of meaning that we give objects is even broader than what we show here—a metal band can signify a lifelong vow, a chiselled stone can commemorate a lost loved one, and a menorah can reveal one's metaphysical beliefs. We hope that our work is a step towards understanding the rich social nature of the physical world.

5.14 Acknowledgments

We thank Hyowon Gweon and Zenna Tavares for useful comments on a previous version of this manuscript. This work was supported by NSF award BCS-2045778.

Chapter 6

Discussion

The world is often thought of as comprised of either physical or social entities. However, like the examples in Figure 1.2 demonstrate, this division is not always so clear-cut. While a collection of rocks scattered across the ground contains no social information, these same rocks assembled into a stack elicit a variety of social inferences: we can infer that the arrangement contains social information (e.g., that it was intentionally assembled by an agent), we can reconstruct what happened (e.g., we can imagine the agent stacking one rock on top of the other), and we can even infer potential mental states (e.g., we can infer why the agent built it). What computations and representations underlie our capacity to reason about physical objects embedded with social information?

Humans have specialized cognitive capacities to process and reason about physical and social entities—intuitive psychology and intuitive physics. These capacities are online from early childhood, supporting a theory-like, causal understanding of physical objects (Spelke and Kinzler, 2007; Téglás et al., 2011) and agents (Gopnik et al., 1997; Wellman, 2014). It is possible that, when reasoning about physical objects in social contexts, both capacities are engaged. While much is understood about these two capacities, less is understood about how they might interact.

The outputs of these capacities are not the only potential inputs to the social representations of physical objects. Before we even realize, some information is automatically and irresistibly processed by our visual system. Previous work in vision science has revealed that this information goes beyond colors, lines, and contrasts, also capturing social properties such as an agent's gender (Bruce et al., 1993; Brown and Perrett, 1993) and dominance and trustworthiness (Todorov and Duchaine, 2008).

To begin elucidating these representations, I first sought to understand three common social inferences that people can make from physical objects. I proposed that detecting whether an arrangement or object contains social information may be, at least partially, driven by low-level visual features. Here I focused on a specific kind of social information: whether an agent was involved in manipulating the arrangement or object. Given that an agent was involved, I then proposed that combining our commonsense physics and commonsense psychology enables us to reason about the agent from how they manipulated the environment.

6.1 Chapter 2-5 Review

In this thesis, I tackled three interconnected inferences: (1) detecting that an object contains social information, (2) reconstructing what an agent did with the object, and (3) inferring the agent's mental states. I introduced previous work showing how our visual system is specialized for certain types of stimuli, including agents. Then, I hypothesized that this specialization might extend beyond agents and to the traces they leave behind (Chapter 2). I presented a series of experiments testing people's visual search performance for these "traces of agency", and found that people were significantly faster and more accurate than when compared to "non-agentic" arrangements.

After presenting some initial evidence for one way in which people may detect whether an object contains social information, I presented an alternative account: Could people detect whether an object contains social information simply by considering the physical plausibility of it occurring naturally? Here I presented a computational model that generated the probability that an environmental manipulation occcurred through some natural physical process (Chapter 3). Then, I shifted towards using this social information as an assumption: given that we know that an object contains social information (e.g., that an agent was involved in manipulating it), what can we learn about what the agent did and their mental states? In Chapter 3, I delved into the first part of this question. I presented a computational model that combines two foundational cognitive capacities—commonsense psychology and commonsense physics—in order to explain how people can reconstruct the actions of an unseen agent. Predictions from this model, combined with human behavioral data, revealed that this inference can be achieved by combining our understanding of how desires lead to actions with our understanding of how our actions lead to physical traces in the environment.

Once we know that there is an agent to reason about, and we can reconstruct their probable past actions, we can infer their mental states. After reviewing a rich body of experimental and computational work detailing how we can infer an agent's unobservable beliefs and desires from their observable actions, I present another computational model of how objects can be used to communicate (Chapter 4). Like in Chapter 3, the use of "communicative objects" can be supported by a combination of commonsense psychology and commonsense physics. Taken together, these threads of research begin to reveal how even physical objects can be perceived and reasoned about as social entities.

6.2 Open questions

This work is a first step in illuminating the computations and representations that underlie our capacity to reason about social information from the physical world. As such, many open questions remain, spanning a wide range of disciplines within psychology, such as cross-cultural research, developmental science, and comparative cognition. For simplicity, I present these areas of future research as its own sub-section below.

6.2.1 How do intuitive psychology and intuitive physics combine?

According to the theory of core knowledge systems (Spelke, 2003; Spelke and Kinzler, 2007), our core systems for objects and agents are *isolated*—meaning that they do not pass information to other systems—and encapsulated—meaning that they do not have access to the information from another system. These informational constraints are present until children acquire another capacity that enables them to combine representations across any conceptual domains: natural language. From these core systems emerge our intuitive theories, abstract causal models of the world (Wellman, 1992; Carey, 2009; Gerstenberg and Tenenbaum, 2017). Here I focused on phenomena where two intuitive theories, intuitive psychology and intuitive physics, are jointly engaged. How do these intuitive theories readily combine?

One potential hypothesis is an extension of the core knowledge account: These intuitive theories inherit the same informational constraints as our core systems (i.e., they are isolated and encapsulated, but with the acquisition of natural language, humans learn to exchange information between them). This makes the prediction that pre-verbal children should be unable to reason about physical objects embedded with social information. A second potential hypothesis is that these intuitive theories do not inherit the same informational constraints as our core systems. This makes the prediction that even pre-verbal children can leverage physical information for social reasoning and vice versa. Finally, a third potential hypothesis, related to the second, is that these intuitive theories contain informational redundancy. That is, our intuitive psychology is theorized to handle agent actions, but may also have a specialized ability to handle agent-object interactions (e.g., similar to the cookie crumb vignette in Chapter 4). This makes similar predictions to the second hypothesis, but makes narrower claims on the cognitive faculties that are engaged.

6.2.2 Is this capacity uniquely human?

If these intuitive theories are isolated and encapsulated, then non-human animals cannot reason about physical objects in a social way (even if they possess both of these intuitive theories) as they lack natural language, the capacity that supports the ability to exchange information across these systems. However, if this hypothesis is false, there may exist another way for information to flow between these systems. One approach to tackling this hypothesis—and illuminating whether this capacity is uniquely human—is to first understand the limits of an animal's social and physical reasoning and then use this to guide predictions about the kinds of social inferences they should be able to make from physical objects.

Previous work done with non-human primates has shown that they are able to reason about physical objects and interactions between them, in a way that resembles human infants (Santos, 2004). In particular, non-human primates can track objects that move behind an occluder (Hauser et al., 1996; Uller et al., 2001), understand that a solid object cannot move through another object (Santos and Hauser, 2002), and that objects cannot move by themselves (Hauser, 1998). Moreover, other work has shown that non-human primates can also represent what others can and cannot see and hear (Santos et al., 2006), as well as what they know (Kaminski et al., 2008; Marticorena et al., 2011). These findings suggest that non-human primates may have the cognitive ingredients to reason about social objects.

Despite the striking similarities between the cognitive capacities of human infants and non-human primates, however, there are also significant functional differences between them, specifically regarding their ability to reason about others. Non-human primates have not been shown to have the capacity to understand that agents can have distinct representations of the world that are decoupled from reality (Martin and Santos, 2016). This stems from previous studies showing that, while non-human primates can represent others' knowledge and ignorance, they lack the kind of belief representation required to pass the false-belief task (Kaminski et al., 2008; Marticorena et al., 2011). This limitation of representational capacity hinders the kinds of social inferences that non-human primates could make from the physical world, as some of these inferences can be quite rich (e.g., reconstructing an agent's past actions; Chapter 4).

Another example of these differences in social reasoning comes from research with vervet monkeys. Vervet monkeys have an astonishing degree of social intelligence, including a nuanced repertoire of vocal calls to signal different types of predators, each associated with different escape responses (Seyfarth et al., 1980a,b). Despite this, vervet monkeys routinely fail to identify predators despite clear physical evidence of their presence. For instance, vervet monkeys fail to infer that a python is hiding in a nearby bush when they encounter the distinct tracks that they leave behind. Similarly, vervet monkeys also fail to infer the presence of a leopard upon encountering a gazelle carcass on a tree (where leopards usually drag their prey so they can feed in solitude; Cheney and Seyfarth, 1985). Critically, this failure appears to persist even after vervet monkeys have, in past events, seen the direct association between the physical evidence and the predator (Cheney and Seyfarth, 1985, 2008). These findings suggest that, despite having the cognitive ingredients for this capacity, we have not yet found a non-human animal that can combine them.

6.2.3 Context in culture

It is no surprise that cultures around the world communicate differently. One key feature that distinguishes the communicative style of a particular culture is how dependent its constituents are on context (Hall, 1976). In high context cultures, much of the information that agents intend to communicate is done so implicitly or subtlety. These cultures tend to be relatively homogeneous and collectivist. On the other hand, low context cultures require communication to be much more direct. These cultures tend to have more diverse populations and be largely individualistic. These distinctions are not a binary one, and cultures can and do form a continuum of context dependence.

These variations between communicative styles may imply that people from each culture may also reason differently about social information within physical objects. However, my proposal relied on foundational cognitive capacities that support social and physical reasoning, so there should not be any differences between people from different cultures in terms of capacity. In Chapter 5, I tested participants from the Tsimane'—a farming-foraging tribe indigenous to the Bolivian Amazon who have limited access to market-integrated communities—and showed that they made similar communicative inferences about the meaning of objects as our U.S. participants.

In terms of usage, however, differences between cultures may exist. For instance, people from any culture should be able to reason about the objects in Chapter 5, but perhaps people from high context cultures use these objects more often, or are faster at making the inference (due to conventionalization; see Section 6.2.5). This could be tested in a handful of ways. Participants could be shown

6.2.4 Parsimony in explanation: heuristics versus Bayesian inference

The computational frameworks I proposed involve relatively complex computations. A seemingly parsimonious alternative is to consider whether heuristics could explain some or all of these findings. Heuristics consists of simple rules that guide agents towards fast solutions to complex problems. However, as I demonstrate below, the critical problem with heuristics is that they generalize poorly, becoming the less parsimonious than the models I proposed.

In Chapter 3, I proposed a computational model that formalized the detection of agents as Bayesian inference over a generative model of natural physical outcomes. In particular, the generative model was built using a physics engine, which both modeled all of the objects in the scene and simulated thousands of physical outcomes. Using Bayesian inference, I computed the probability that a particular scene occurred naturally and used that to further compute the probability that an agent was involved. However, a simpler approach that participants may have been using is to rely on two key features of the stimuli: how far the blocks were placed from the funnel and whether they were stacked. This may explain a great deal of the variance in participant judgments that my model did. However, one example where these heuristics would fail is if a circle made of blocks was arranged close to the funnel. According to the heuristics approach, participants should not find this to be the work of an agent, despite the peculiarity of this pattern occurring by chance. While this is also something not yet handled by my model, I hope to pursue this in future work.

In Chapter 4, I presented a computational model that formalized mental event reconstruction as Bayesian inference over a generative model of environmental traces. Here I directly accounted for the possibility of heuristics in the design of the stimuli (e.g., ensuring that the target goal was not always the one closest to the cookie crumbs, and that it could not be determined by projecting a straight line that intersected the entrance and the location of the cookie crumbs). Additionally, I computed two multinomial logistic regressions trained to predict participant judgments as a function of the distance between the pile of cookie crumbs and each goal, the average distance between the pile of cookie crumbs and each door, the number of doors, and all of their interactions (Experiment 3 also had an additional predictor for the number of agents in the room). Both of these regressions explained significantly less variance in participant data than my proposed model.

Finally, in Chapter 5, I presented a computational model that formalized communication through objects as recursive Bayesian inference. Here I directly addressed the heuristics alternative throughout by addressing conventions. Conventions, which perform the same functional role as heuristics, are a key part of how I hypothesize that communicative objects become widespread. However, in the experiments I presented, I tackle the question of how people navigate around objects that do not have a conventional meaning.

Despite the initial parsimony of heuristics-based approaches, they tend to lack the ability to generalize and, when employed in the experiments here, failed to capture the fine-grained patterns of participant judgments. On the other hand, the models I proposed not only explained graded participant judgments with quantitative accuracy, but can also be easily extended to account for broader kinds of inferences. While the work I presented here posed a problem for heuristics-based approaches, it remains an open question whether there exist heuristics that are flexible enough to capture the inferences I studied.

6.2.5 Efficient computation and amortized inference

Heuristics tend to be an appealing alternative for scientific parsimony, but also for computational efficiency. Bayesian inference, especially in problems with very large hypothesis spaces, can be computationally expensive or intractable. Previous work has shown that some inferences that were previously thought to be handled by high-level cognition (e.g., our intuitive theories), are actually “baked into” low-level visual processing. For instance, Battaglia et al. (2013) propose an intuitive physics engine framework to explain how people make stability judgments, however more recent work has shown that this capacity might be supported by more basic visual processes

(Firestone and Scholl, 2016).

In addition to built-in computations, visual processing can also support when these computations occur. In Chapter 2, I proposed that low-level visual features may serve as a filter for social reasoning. That is, instead of applying Theory of Mind over every object in our visual experience—which would be computationally demanding—our visual system handles when an object should be reasoned about further.

Some computations are not handled by low-level visual processing in any significant way, yet still improve in efficiency over repeated interactions. For instance, in Chapter 5, I discussed the phenomena of communicative objects and how, often in our everyday life, interpreting the meaning of these objects seems incredibly trivial. I attributed this triviality to convention and pedagogy, and argued that these are the tools by which the meaning of communicative objects becomes widespread. Conventions, in particular, may be a way that humans perform *amortized inference*, where the goal is to figure out how to re-use inferences in future situations (Gershman and Goodman, 2014).

In the series of accounts that I presented, another question that remains is to what extent are these inferences automatic and irresistible, rather than only purposefully engaged. In the cue-based account of Chapter 2, my proposal is that these computations are happening at the level of visual processing. While it is currently not yet known whether the visual system is handling the recognition of social information (versus simply recognizing small errors), this level of processing is automatic. On the other hand, in Chapters 3, 4, and 5 I tested people on their ability to perform an explicit inference. These kinds of inferences are necessary for navigating the environment, lending credibility to the idea that they, too, may be automatically engaged. However, it remains to be tested whether participants would perform this kind of reasoning when their current task is irrelevant to the inference.

6.2.6 When and how do agents come into the picture?

In Chapter 2, I presented a cue-based account of how people come to detect that an object contains social information. I presented two visual search experiments that revealed that people have specialized attention for slightly-misaligned block towers over perfectly-aligned ones. One possibility is that people are perceiving *agentiveness.* That is, that the visual system interprets the slightly-misaligned block towers as containing social information. In face perception, it was previously thought that certain high-level features, like trustworthiness, could only be inferred after inferring the identity of the face. However, previous work has shown that to not be the case, instead revealing that these judgments occur simultaneously (Todorov and Duchaine, 2008). Moreover, recent work has shown that our visual system intrinsically computes basic social relations (Isik et al., 2017). This hypothesis makes the prediction that these inferences should show activation only in visual cortex.

An alternative possibility is that our visual system is instead sensitive to small errors, leaving high-level cognition to handle the detection of agents. This resembles how vision processes an agent's movement, but high-level cognition interprets the agent's underlying mental states. This hypothesis makes the prediction that these inferences may also show activation in brain regions other than visual cortex, such as in the medial prefrontal cortex or temporo-parietal junction, two regions known to be engaged during social reasoning (Saxe and Powell, 2006). Investigating these predictions could be tested using fMRI.

6.2.7 Inferring mental states beyond communicative intent

In Chapter 5, I proposed a theoretical account of how people infer others' mental states from physical objects. I focused on communicative intent as a case study, investigating how objects can reveal whether an agent should approach or avoid something in the environment. While my account explained the fine-grained patterns of participant judgments, the scope of this work was limited to a small subset of communicative meanings and an even smaller subset of mental states.

Beyond serving as deterrent or invitational signals, objects can also communicate one's moral and metaphysical beliefs (e.g., through religious symbols) or wealth (e.g., through lavish jewelry). Furthermore, objects can reveal how long ago something was built, how much effort it required or how many people were required to build it, how much care was put into it, and even to what extent the goals were functional or aesthetic. I hypothesize that all of these inferences are also supported by a combination of our intuitive psychology and intuitive physics, a more complex combination than I proposed here, but this remains an open question.

6.2.8 Detecting intentional action

Without knowing whether an object was placed intentionally, we would clash in all sorts of social situations, such as not realizing that an object could indicate that a parking spot is taken (e.g., Figure 1.2i) or that a certain path should be avoided (e.g., Figure 1.2l). Even infants have an understanding of intentional, goal-directed actions (Gergely et al., 1995; Woodward and Sommerville, 2000). However, since agents and their actions are unobservable in the problem I consider here, how do we know whether an arrangement or object has been intentionally manipulated, as opposed to being the result of a natural force (e.g., the wind) or even the result of agent *accidentally* interacting with it? And to what extent is the solution to this problem divided between perception and cognition?

Arrangements of objects that appear intentional often do because there is a near-zero chance that the environment could produce them. Consider a cairn: what are the chances that a stack of rocks, say two feet in height, naturally occurs in a forest? Even through simple combinatorics, only a handful of configurations end up making an upright stack. This inference seems to be particularly sensitive to the environment. If the rocks lie at the base of a cliff, this inference suddenly flips, since piling would be a natural occurrence. If placed in the middle of one of two hiking trails, it swings back in the other direction, not only because of the implausibility of occurring naturally, but also because we can imagine a reason that an agent would have built it there.

Having prior expectations about what agents desire also gives us an expectation over their probable actions. If valid proposals to explain an action fail to come to mind, then we may label the action as an accident. The cairn, for instance, would be difficult to interpret if, instead of being placed along the middle of one of two available hiking paths, was placed far away from the trails, in the middle of the forest (here it might even be inferred to be art). A more complete account of how we represent physical objects containing social information must also include an explanation of how we disentangle intentional, accidental, and natural actions.

6.3 Conclusion

Taken together, the work I have presented here attempts to advance our understanding of human social cognition. First, I presented an account that suggests that the visual system is specialized to process social information from physical objects that reveals the involvement of an agent. Second, I presented a computational model that shows that the combination of two foundational cognitive capacities can explain how we reconstruct an agent's past actions. Finally, I presented another computational model that similarly combines these two capacities, this time showing how they can support the use of objects to convey social information, like mental states. This work contributes to our understanding of human social intelligence transforms our experience of the physical world.

Bibliography

Aguiar, A. and Baillargeon, R. (1999). 2.5-month-old infants' reasoning about when objects should and should not be occluded. *Cognitive psychology*, 39(2):116–157.

Aquinas, T. (1485). *Summa Theologica*.

Arechar, A. A. and Rand, D. G. (2021). Turking in the time of covid. *Behavior research methods*, 53(6):2591–2595.

Back, E. and Apperly, I. A. (2010). Two sources of evidence on the non-automaticity of true and false belief ascription. *Cognition*, 115(1):54–70.

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, 2(3):179–200.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064.

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

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

Becker, F. D. (1973). Study of spatial markers. *Journal of Personality and Social Psychology*, 26(3):439.

Becker, F. D. and Mayo, C. (1971). Measurement effects in studying reactions to spatial invasions.

Bellman, R. (1957). A markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684.

Bennett-Pierre, G., Asaba, M., and Gweon, H. (2018). Preschoolers consider expected task difficulty to decide what to do and whom to help. In *Cogsci*.

Birch, S. A. and Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5):382–386.

Bonawitz, E., Denison, S., Gopnik, A., and Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, 74:35–65.

Bonda, E., Petrides, M., Ostry, D., and Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, 16(11):3737–3744.

Bridgers, S., Jara-Ettinger, J., and Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 4(2):144–152.

Brown, E. and Perrett, D. I. (1993). What gives a face its gender? *Perception*, 22(7):829–840.

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., and Linney, A. (1993). Sex discrimination: how do we tell the difference between male and female faces? *perception*, 22(2):131–152.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Cheney, D. L. and Seyfarth, R. M. (1985). Social and non-social knowledge in vervet monkeys. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308(1135):187–201.

Cheney, D. L. and Seyfarth, R. M. (2008). *Baboon metaphysics: The evolution of a social mind*. University of Chicago Press.

Cholewiak, S. A., Fleming, R. W., and Singh, M. (2013). Visual perception of the physical stability of asymmetric three-dimensional objects. *Journal of vision*, 13(4):12–12.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, 49(6):1407–1409.

Collette, S., Pauli, W. M., Bossaerts, P., and O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, 6:e29718.

Colombatto, C. and Scholl, B. J. (2022). Unconscious pupillometry: An effect of "attentional contagion" in the absence of visual awareness. *Journal of experimental psychology: general*, 151(2):302.

Conway, B. R., Ratnasingam, S., Jara-Ettinger, J., Futrell, R., and Gibson, E. (2020). Communication efficiency of color naming across languages provides a new framework for the evolution of color terms. *Cognition*, 195:104086.

Dawkins, R. et al. (1996). *The blind watchmaker: Why the evidence of evolution reveals a universe without design.* WW Norton & Company.

Dennett, D. C. (1989). *The intentional stance.* MIT press.

Devaine, M., Hollard, G., and Daunizeau, J. (2014). The social bayesian brain: does mentalizing make a difference when we learn? *PLoS computational biology*, 10(12):e1003992.

Di Giorgio, E., Lunghi, M., Simion, F., and Vallortigara, G. (2017). Visual cues of motion that trigger animacy perception at birth: The case of self-propulsion. *Developmental science*, 20(4):e12394.

Edney, J. J. and Jordan-Edney, N. L. (1974). Territorial spacing on a beach. *Sociometry*, pages 92–104.

Ericson, S. R., Denison, S., Turri, J., and Friedman, O. (2023). Probability and intentional action. *Cognitive Psychology*, 141:101551.

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

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., and Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, 113(34):E5072–E5081.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition.*

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1):3–44.

Gao, T., McCarthy, G., and Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science*, 21(12):1845–1853.

Gergely, G. (2003). The development of teleological versus mentalizing observational learning strategies in infancy. *Bulletin of the Menninger clinic*, 67(2: Special Issue):113–131.

Gergely, G. and Csibra, G. (1997). Teleological reasoning in infancy: The infant's naive theory of rational action: A reply to premack and premack. *Cognition*, 63(2):227–233.

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

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

Gershman, S. and Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36.

Gerstenberg, T., Goodman, N., Lagnado, D., and Tenenbaum, J. (2020). A counterfactual simulation model of causal judgment.

Gerstenberg, T. and Tenenbaum, J. B. (2017). Intuitive theories.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S. T., and Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Gopnik, A., Meltzoff, A. N., and Bryant, P. (1997). Words, thoughts, and theories.

Gosling, S. D., Gaddis, S., and Vazire, S. (2008). First impressions based on the environments we create and inhabit. *First Impressions*, pages 334–356.

Gosling, S. D., Ko, S. J., Mannarelli, T., and Morris, M. E. (2002). A room with a cue: personality judgments based on offices and bedrooms. *Journal of personality and social psychology*, 82(3):379.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., and Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of cognitive neuroscience*, 12(5):711–720.

Grossman, E. D. and Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, 35(6):1167–1175.

Grossman, E. D., Blake, R., and Kim, C.-Y. (2004). Learning to see biological motion: Brain activity parallels behavior. *Journal of Cognitive Neuroscience*, 16(9):1669–1679.

Gweon, H., Asaba, M., and Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. In *CogSci*.

Hall, E. T. (1976). *Beyond culture.* Anchor.

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

Hauser, M. D. (1998). A nonhuman primate's expectations about object motion and destination: The importance of self-propelled movement and animacy. *Developmental Science*, 1(1):31–37.

Hauser, M. D., MacNeilage, P., and Ware, M. (1996). Numerical representations in primates. *Proceedings of the National Academy of Sciences*, 93(4):1514–1517.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., and Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, 317(5843):1360–1366.

Ho, M. K., Cushman, F. A., Littman, M., and Austerweil, J. L. (2019). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General.*

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F. A., and Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, pages 3027–3035.

Hurwitz, E., Brady, T., and Schachner, A. (2019). Detecting social transmission in the design of artifacts via inverse planning.

Hurwitz, E. and Schachner, A. (2020). People use inverse planning to rationally seek social information from objects. In *CogSci.*

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152.

Jacobs, C., Lopez-Brau, M., and Jara-Ettinger, J. (2021). What happened here? children integrate physical reasoning to infer actions from indirect evidence. *CogSci.*

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110.

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

Jara-Ettinger, J., Piantadosi, S., Spelke, E. S., Levy, R., and Gibson, E. (2017). Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental Science*, 20(6):e12459.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. (2020a). The naïve utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. (2020b). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334.

Jern, A. and Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, 142:12–38.

Jern, A., Lucas, C., and Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. *Advances in Neural Information Processing Systems*, 24:2276–2284.

Jern, A., Lucas, C. G., and Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211.

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological research*, 38(4):379–393.

Johnson, M. H., Dziurawiec, S., Ellis, H., and Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2):1–19.

Kaminski, J., Call, J., and Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109(2):224–234.

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

Keil, F. C. and Newman, G. E. (2015a). Order, order everywhere, and only an agent to think: The cognitive compulsion to infer intentional agents. *Mind & Language*, 30(2):117–139.

Keil, F. C. and Newman, G. E. (2015b). Order, order everywhere, and only an agent to think: The cognitive compulsion to infer intentional agents. *Mind & Language*, 30(2):117–139.

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Kool, W. and Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(12):899–908.

Kushnir, T., Xu, F., and Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological science*, 21(8):1134–1140.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40.

Leslie, A. M. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288.

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

Lieder, F. and Griffiths, T. L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:1–85.

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

Lopez-Brau, M. (2021). Social inferences from physical evidence.

Lopez-Brau, M., Colombatto, C., Jara-Ettinger, J., and Scholl, B. (2021). Attentional prioritization for historical traces of agency. *Journal of Vision*, 21(9):2748–2748.

Lopez-Brau, M. and Jara-Ettinger, J. (2020). Physical pragmatics: Inferring the social meaning of objects.

Lopez-Brau, M., Kwon, J., and Jara-Ettinger, J. (2022). Social inferences from physical evidence via bayesian event reconstruction. *Journal of Experimental Psychology: General*.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., and Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3):e92160.

Ma, L. and Xu, F. (2013). Preverbal infants infer intentional agents from the perception of regularity. *Developmental psychology*, 49(7):1330.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Marticorena, D. C., Ruiz, A. M., Mukerji, C., Goddu, A., and Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental science*, 14(6):1406–1416.

Martin, A. and Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in cognitive sciences*, 20(5):375–382.

Matsumoto, D. and Yoo, S. H. (2006). Toward a new generation of cross-cultural research. *Perspectives on psychological science*, 1(3):234–250.

McDermott, J. H., Schultz, A. F., Undurraga, E. A., and Godoy, R. A. (2016). Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature*, 535(7613):547–550.

Misyak, J. B. and Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130487.

Misyak, J. B., Melkonyan, T., Zeitoun, H., and Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*, 18(10):512–519.

Misyak, J. B., Noguchi, T., and Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science*, 27(12):1550–1561.

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603.

Newman, G. E., Keil, F. C., Kuhlmeier, V. A., and Wynn, K. (2010). Early understandings of the link between agents and order. *Proceedings of the National Academy of Sciences*, 107(40):17140–17145.

Oey, L. A., Schachner, A., and Vul, E. (2022). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General.*

Onishi, K. H. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, 308(5719):255–258.

Oring, E. (1986). *Folk groups and folklore genres: An introduction.* University Press of Colorado.

Pelz, M., Schulz, L., and Jara-Ettinger, J. (2020). The signature of all things: Children infer knowledge states from static images. *PsyArXiv.*

Pesowski, M. L., Kelemen, D., and Schachner, A. (2021). Children use artifacts to infer others' shared interests. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Pesowski, M. L., Quy, A. D., Lee, M., and Schachner, A. (2020). Children use inverse planning to detect social transmission in design of artifacts. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Peuskens, H., Vanrie, J., Verfaillie, K., and Orban, G. A. (2005). Specificity of regions processing biological motion. *European Journal of Neuroscience*, 21(10):2864–2875.

Piantadosi, S. T., Jara-Ettinger, J., and Gibson, E. (2014). Children's learning of number words in an indigenous farming-foraging group. *Developmental Science*, 17(4):553–563.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International journal of psychology*, 24(6):737–756.

Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, 17(5):1215–1233.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, 36(1):1–16.

Quillien, T. and German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, 214:104806.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological science*, 27(9):1192–1206.

Repacholi, B. M. and Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1):12.

Rossano, F., Fiedler, L., and Tomasello, M. (2015). Preschoolers' understanding of the role of communication and cooperation in establishing property rights. *Developmental Psychology*, 51(2):176.

Royka, A., Chen, A., Aboody, R., Huanca, T., and Jara-Ettinger, J. (2022). People infer communicative action through an expectation for efficient communication. *Nature Communications*.

Santos, L. R. (2004). 'core knowledges': a dissociation between spatiotemporal knowledge and contact-mechanics in a non-human primate?

Santos, L. R. and Hauser, M. D. (2002). A non-human primate's understanding of solidity: dissociations between seeing and acting. *Developmental Science*, 5(2):F1–F7.

Santos, L. R., Nissen, A. G., and Ferrugia, J. A. (2006). Rhesus monkeys, macaca mulatta, know what others can and cannot hear. *Animal Behaviour*, 71(5):1175–1181.

Sarin, A., Ho, M. K., Martin, J. W., and Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, 208:104544.

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

Saxe, R. and Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699.

Saxe, R., Tenenbaum, J., and Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, 16(12):995–1001.

Schachner, A., Brady, T., Oro, K., and Lee, M. (2018). Intuitive archeology: Detecting social transmission in the design of artifacts.

Schachner, A. and Kim, M. (2018). Alternative causal explanations for order break the link between order and agents.

Scholl, B. J. and Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment. *Social perception: Detection and interpretation of animacy, agency, and intention*, 4629.

Scontras, G., Tessler, M. H., and Franke, M. (2018). Probabilistic language understanding: An introduction to the rational speech act framework. *URL https://gscontras. github. io/probLang.*

Scott, R. M. and Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, 21(4):237–249.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980a). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, 210(4471):801–803.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980b). Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Animal Behaviour*, 28(4):1070–1094.

Shaffer, D. R. and Sadowski, C. (1975). This table is mine: Respect for marked barroom tables as a function of gender of spatial marker and desirability of locale. *Sociometry*, pages 408–419.

Shafto, P., Goodman, N. D., and Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., and Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40:99–124.

Simion, F., Macchi Cassia, V., Turati, C., and Valenza, E. (2001). The origins of face perception: specific versus non-specific mechanisms. *Infant and Child Development: An International Journal of Research and Practice*, 10(1-2):59–65.

Sommer, R. and Becker, F. D. (1969). Territorial defense and the good neighbor. *Journal of personality and social psychology*, 11(2):85.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, 14(1):29–56.

Spelke, E. S. (2003). What makes us smart? core knowledge and natural language. *Language in mind: Advances in the study of language and thought*, pages 277–311.

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

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

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Todorov, A. and Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive neuropsychology*, 25(3):395–410.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual cognition*, 14(4-8):411–443.

Troje, N. F. (2013). What is biological motion? definition, stimuli and paradigms. *Social perception: Detection and interpretation of animacy, agency, and intention*, pages 13–36.

Turk-Browne, N. B., Jungé, J. A., and Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, 134(4):552.

Uller, C., Hauser, M., and Carey, S. (2001). Spontaneous representation of number in cotton-top tamarins (saguinus oedipus). *Journal of Comparative Psychology*, 115(3):248.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., and Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems*, pages 1874–1882.

van Buren, B., Uddenberg, S., and Scholl, B. J. (2016). The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic bulletin & review*, 23(3):797–802.

Van de Vijver, F. J. and Leung, K. (2021). *Methods and data analysis for cross-cultural research*, volume 116. Cambridge University Press.

Von Grünau, M. and Anston, C. (1995). The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, 24(11):1297–1313.

Warneken, F. and Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *science*, 311(5765):1301–1303.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences.* Rand McNally Chicago.

Wellman, H. M. (1992). *The child's theory of mind.* The MIT Press.

Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford University Press.

Wellman, H. M. and Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition*, 30(3):239–277.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34.

Woodward, A. L. and Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, 11(1):73–77.

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J., and Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk.