Abstract
How Do I Know What You Know? A Novel Theoretical Account of Epistemic Inference
Rosie Aboody
2022

Of the capacities that make us uniquely human—pedagogy, social learning, cooperation, communication, moral evaluation—all hinge, at least in part, on an understanding of what others know or believe. Critically, we cannot see mental states: we have access only to the observable behaviors they cause. So, to navigate the social world, we must often infer what others think from observing what they do. While prior work has investigated how children and adults infer preferences, goals, and desires from behavior, little research has investigated how we infer epistemic states (knowledge and beliefs). In this thesis, I expand upon existing accounts of mental state reasoning to formalize a novel theoretical account of epistemic inference. I test whether this account captures adults' knowledge inferences, and use it to systematically investigate the development of our capacities.

How Do I Know What You Know? A Novel Theoretical Account of Epistemic Inference

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

By
Rosie Aboody

Dissertation Director: Julian Jara-Ettinger

May 2022

This dissertation is dedicated to my family, who have always supported me; to the friends who kept me sane; and to the teachers and mentors who guided me.

Acknowledgments

First and foremost, I want to thank my family, for all your love and support even when I'm a pain. I could never have done this without you.

I want to thank Julian for being the best advisor I could have ever had the privilege of working with. There are really no words for how grateful I am to have been your first graduate student. You've taught me not only how to be a good scientist, but also how to be a kind and thoughtful mentor and colleague. Thank you.

I want to thank Frank for being the most supportive mentor I know. The privilege of being a part of your lab has enriched my time here beyond measure. Thank you for the panda costumes, goat visits, and beautiful science.

I want to thank Laurie for being one of the sharpest empiricists I've ever met, and for supporting me even when my interests turned more towards development. Thank you for Howloween, the Silliman community, and for teaching the best course I've ever taken (which has shaped the way I think about pretty much everything related to Theory of Mind).

Going back to the beginning, I want to thank the members of the Gopnik Lab, for creating an exceptional environment where students learn by doing. Alison: thank you for being brilliant, and helping open the door a little further for all the women scientists who followed. Caren: thank you for being an amazing first academic mentor. As I've moved through graduate school myself, it's become more and more obvious just how exceptional you were (and how hard you must have worked to support us mentees). Thank you for taking me on, and always supporting me (even when I was interested in Theory of Mind—yuck!) Sophie: thank you for being the best "academic big sister" I could've asked for, and for being endlessly supportive throughout my years in the lab. Azzu, Ny, Zac, Adrienne, Shaun, Mariel, and Sophia: thank you for being amazing lab members. Our dev community was so vibrant and fun, and I will forever miss our games nights and Gregoire's lunches. And a special thanks to Mariel, for always being herself (with all the wonders that entails).

I want to thank the members of the Comparative Cognition Lab, my first home at Yale. Thank you Laurie again, for bringing me here and supporting me no matter what. Thank you Angie, Mikey, Alyssa, Chelsey, Lindsey, Matt, and of course, Vader. You created a community so filled with warmth and joy in science, and I feel privileged to have been allowed in.

A special thanks to Angie and Mark: you two really took me under your wing,

both in the lab and out. Thank you for all the games nights, trips (what would my life be like if you didn't take me to places like Book Barn or Bloodroot?), Vader walks, advice sessions, and Claire's lunches. You two made me feel at home, and taught me so much of what makes me the person and scientist I am today.

A special thanks also to Alyssa: you are one of the most thoughtful, emphatic, and kind people I've ever had the privilege of befriending. You made my time here so much better.

I want to thank my cohort, who have been here with me through it all. Sami: thank you for being my friend from the beginning. All of those coffee runs, Hillhouse strolls, Claire's dinners, and ultimately, climbing sessions and baking evenings, made my life so much better. I don't know that I would have survived my first year without you. Thank you Shina, Estée, Paola, Brian Earp, Qi, Natalie, Julian, Lauren, Erica, Kristen, Emily C., and Brian Bink for accompanying me on this journey. And thank you Nathan for being my bonus bud throughout my first year. All of you enriched my time here, and made it 1000x more enjoyable.

I also want to thank the many friends I made in other cohorts and departments who made my time here a pleasure. Sky: thank you for many things, including but not limited to olive vegetables, gai lan, and a multitude of other foods; for belaying when I'm not injured and making me go to the doctor when I am; for never making me navigate; and for infinite patience. Emily and Daniel: thank you for becoming some of my best friends; for LOTR day, for the Star Wars nights, Passover mash-ups, Hannukah sweaters, raspberry picking, and retreats; for being my pandemic pod and becoming my family away from home. Gracie: thank you for bringing Manny into our lives; for cat and cheese; for Halloween parties and scaring the crap out of me with Black Mirror; and for all the rest. Flora and Winston: thank you climbing, for always cooking me delicious food, for goat screams, baking, and so much more. Ryan: thank you for being one of my best friends here; for thrifting and protein pucks and East Rock walks and Book Barn trips and Bay Area reminiscing. Vanessa: thank you for being one of the coolest people I know; for making the best damn hummus in town; for Hobbes, for Hen & Heifer, and for keeping me sane during my time here (but especially for Hen & Heifer). Alec: thank you for accompanying me through some of the most challenging times in my program, and for reminding me that I actually really like things like cooking, live music, and plays. April, Bud, Clara, and Sam: thank you for being the senior Yalies who took us young ones under your wing. Our summer evenings at the Gryphon brought me some of the first sense of community I felt after moving to the East coast (not to mention the best sea shanties in town). Ashley, Julia, Kate, Zach, Sifana, and Pinar: thank you for being an amazing dev community. I couldn't imagine a more supportive crew. Ryan, Vlad, Ajua, and Judy: thank you for being an amazing soc community, for listening to my many rants and stressors, and always being the most kind and thoughtful crowd. Shina and Cole: thank you for wine & whine, for park sits, evening walks, and friendship. Tim: thank you for museum strolls, Atticus hangs, and helping ensure my pandemic cooking never

went to waste. Kat, Estée, Megha and Lucinda: thank you for heading DivComm in recent years (and thanks to the many coordinators before that). You have served such a crucial (and unpaid) role in our department, and have made it a much better place. Enya and Ming: thank you for being amazing collaborators on the pipeline initiative (and thank you Bennett for starting the initiative in the first place!) It's been such a joy to work with you.

I also want to thank my "conference crew" spread across institutions: Mika, Lonnie, Liz, Emily L., Joe, Junyi, Holly, Ben, Claire, Carolyn, and many, many more. You are a large part of what makes academia so fun. Thank you for table-tots, donuts, poutine, good conversation, and the most concentrated learning I do on a yearly basis.

Moving along semi-chronologically, I want to thank the members of the Computational Social Cognition Lab. Madison: thank you for always witnessing the shenanigans with me. Michael: thank you for being our lab's international (national?) man of mystery, and for taking our pranks with good humor. Amanda: thank you for all the quiet moments of magic you create, and for your unending enthusiasm for the beauty and acoustics of the Swiss Alps (Dr. Belthoffer would be so proud). Colin: my soul-sister, my Rock, my favorite Jake-owner, thank you for bringing Jake into our lives. Oh, and I guess you're pretty great as well. Thank you for all the book barn trips, park hangs (will never forget your air couch), nametags, NY trips, Ikea karaoke, and general prankery. Gina: thank you for being your wonderful self, and bringing so much spicy memery to our lab. And for becoming the first Official Science Human!! An example for us all. Marlene: thank you for always being a merciful goddess, plant parent, intrepid adventurer, and lab fire-starter. Given my love for fire, you can guess how important a role you have played in my time here. Mackenzie: Thank you for keeping the lab together. Isaac: thank you for being the coolest possible postdoc, and owning such a wonderful dog. And for being an amazing collaborator! Daniel: thank you for being the most agreeable person I've probably ever met. I needed someone to balance out my distinct lack thereof... too bad you took 5 years to show up. Mika: I'm so excited that after 7 years, we finally achieved our goal of being labmates! Thank you for accompanying me on pretty much my entire academic journey, and making it so much better. And finally, Julian: there is too much to list, but thank you again, for everything. Perhaps especially, for being a good sport throughout all the pranks, shenanigans, and unsolved sock-related mysteries. Grad school is hard, but our lab made it fun.

I would also like to thank the members of the Cognition and Development Lab. Angie, Mark, Matt, Sam, and Rick: thank you for being the best senior lab members I could've asked for. Thank you for your mentorship, advice, and help throughout the years. Aaron: thank you for all the tea, conversations, and shenanigans. Alexander: thank you for the good ideas, good company, and goats. Emory, Mandy, Sarah, Maureen, Sara and Nicole: thank you for being wonderful labmates, and for good times with the creepy doll. Flora: thank you for Emma, for the plants, for the chocolate, for everything, and most importantly, for just being you. Sami: thank you

for holding the lab together, and for being wonderful. And finally, thank you Frank, for being a constant source of joy during the long grad school process.

I would also like to thank the members of our newest dev lab, the Leonard Learning Lab: Brandon, Reut, Melissa, and Mika. Despite the short time we have overlapped, you have brought me great joy, and many snacks (which have also contributed to the joy). Thank you Julia, for being a brilliant scientist and an amazing mentor (I loved TAing your class)!

I would also like to thank the amazing Psych admin team, past and present: Lynn, Patty, Andrea, Christine, Krystal, Varvara, Jim, Bret, Kim, Kelly, and Sam. Thank you for keeping the place running (literally), and making my time here possible.

Finally, an enormous thanks to all those I have failed to mention in this rambling section. I confess—the acknowledgements seemed like the least urgent part of the thesis, and now here we are, at 2am on the day my thesis must be submitted to the university. So, although I could go on and on all night, I will stop here by simply thanking all of those who have helped me along my path, accompanied me on it, and improved my life.

Contents

Chapter 1

Introduction

Humans are unique in the extent to which we rely on social cognition to navigate the world: we rely on others to teach us more than we could ever discover on our own, pass on our knowledge in turn, and coordinate to accomplish feats none of us could have accomplished alone. For decades, psychologists have worked to uncover the foundations and development of social cognition, yielding a productive debate about the nature of our understanding of other minds. While some researchers have suggested that many aspects of human social cognition are innate (e.g., Carruthers, 2002; Spelke, 2003), other researchers have suggested that much of our social cognition relies upon a theory of how other minds work, which develops throughout the preschool years (called a "Theory of Mind"; e.g., Gopnik & Wellman, 1994; Wellman, 2014). And yet, despite long-standing interest, basic questions about the nature of our capacities remain unresolved. Instead, the research arising from this debate has opened new puzzles about the nature and development of human mental state reasoning.

1.1 An old puzzle: Mental state reasoning is not developmentally cohesive

Past research has sought to identify which aspects of mental state reasoning are reliably early-emerging. This work finds that from the first years of life, children can represent and infer others' goals (Gergely & Csibra, 2003; Woodward, 1998), intentions (Meltzoff, 1995) and desires (S. Liu, Ullman, Tenenbaum, & Spelke, 2017; Repacholi & Gopnik, 1997). By preschool, children readily reason explicitly about mental states like desires (see Wellman & Liu, 2004), and react appropriately: protesting intentional harms more than honest mistakes (Josephs, Kushnir, Gräfenhain, & Rakoczy, 2016; see also Kachel, Svetlova, & Tomasello, 2018), inferring what others like by observing the costs they're willing to incur (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), and deciding what actions to imitate by considering the goals an agent is pursuing (Bekkering, Wohlschlager, &

Gattis, 2000). Thus, an understanding of mental states like goals and desires appears early-emerging and robust.

In contrast to children's precocious goal- and desire-reasoning, a full ability to understand and reason about others' beliefs and knowledge may not be in place until the end of preschool. Although infants may be able to represent simple false beliefs (e.g., Onishi & Baillargeon, 2005; but see Kampis, Karman, Csibra, Southgate, & Hernik, 2021; Powell, Hobbs, Bardis, Carey, & Saxe, 2018), preschoolers struggle to explicitly do the same (failing to appropriately track others' false beliefs and predict behavior on this basis until age four or five; Wellman, Cross, & Watson, 2001). Furthermore, preschoolers do not reliably consider the reasons behind past errors when deciding whom to trust (Bridgers, Buchsbaum, Seiver, Griffiths, & Gopnik, 2016; Nurmsoo & Robinson, 2009b), struggle to integrate emotional expressions into their understanding of beliefs (Bradmetz & Schneider, 1999; Wu & Schulz, 2018), and do not understand how mental states like ignorance will lead agents to act (Chen, Su, & Wang, 2015; Friedman & Petrashek, 2009). Thus, an explicit understanding of knowledge and belief appears later-developing, and may not be fully robust until the end of preschool.

Why might epistemic reasoning emerge later than goal-, desire-, and intention-reasoning? Some prior research proposes that epistemic reasoning does not emerge later, per se—but rather, an ability to express or use it is what develops (Carlson, Moses, & Hix, 1998; Carlson, Moses, & Breton, 2002; Frye, Zelazo, & Palfai, 1995; Hala & Russell, 2001; Hughes, 1998; Russell, 1996; Zelazo, Carter, Reznick, & Frye, 1997). Indeed, it is likely the case that as children develop more robust executive functions (such as working memory and inhibitory control), many aspects of mental state reasoning become more tractable: for instance, children may better inhibit their own beliefs to reason about those of others. However, while the development of domain-general capacities likely helps children reason about other minds more effectively, this work also leaves open an important question: **does children's Theory of Mind undergo domain-specific conceptual development throughout the preschool years?**

1.2 A new puzzle: Epistemic reasoning is not developmentally cohesive either

While researchers have sought to understand the development of epistemic reasoning for decades, the majority of past research has relied primarily on one task—the false belief task—as a conservative measure of Theory of Mind. This task measures whether children can track others' beliefs, even when those beliefs differ from their own (see Wellman et al., 2001), and is generally regarded as the gold-standard measure of Theory of Mind. Thus, most past research has not sought to investigate the breadth of preschoolers' epistemic abilities. Instead, over the last 35 years, our field's focus

on the false belief task as the measure of Theory of Mind has led many researchers to treat mental state reasoning as a binary. Children or non-human primates who pass the false belief task are said to have a Theory of Mind; those who fail do not (see Tomasello, 2018 for a historical overview).

While the false belief task has served the field in important ways, it sheds light on only a portion of epistemic reasoning. For instance, reasoning over ignorance has long been considered more basic than reasoning over false beliefs—but recent work suggests that preschoolers who pass the false belief task still struggle to understand how ignorant agents will act (Chen et al., 2015). These and related findings suggest that false belief may not be the sole measure of a Theory of Mind; rather, it may be only one of several important abilities needed to fully reason about other minds (see also Phillips et al., 2021). If this is the case, then to understand the development of epistemic reasoning, it may be more productive to think of Theory of Mind not as a binary, but as a continuum: **characterizing the representations and computations required for different mental state reasoning tasks, and investigating how each is acquired.**

1.3 What are the basic components of a Theory of Mind?

To resolve these puzzles, we must first define the basic components of a Theory of Mind, and then investigate how they emerge. Luckily, much classic research in our field has worked to characterizing our understanding of other minds (Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Wellman & Woolley, 1990). Figure 1.1a depicts a simplified version of a classic Theory of Mind model (Wellman, 1992). Broadly, such accounts of Theory of Mind are comprised of two components: an ability to represent mental states, and a causal model that specifies how mental states give rise to behavior. For instance, to pass a false belief task, children must be able to represent an agent's belief, and predict how this belief will lead an agent to act (given their causal understanding of how beliefs and desires combine to produce action; for more, see Wellman, 1992, 2014).

Initially, researchers sought to understand whether young children were capable of representing mental states and using these representations to support downstream reasoning (for example, predicting how mental states will lead to action). This work showed that from the first years of life, children are capable of representing mental states like goals (Gergely & Csibra, 2003; Woodward, 1998), beliefs (Onishi & Baillargeon, 2005), and desires (Repacholi & Gopnik, 1997). Furthermore, young children clearly use mental state representations to predict agents' actions and react appropriately (e.g., reacting with surprise when these predictions are violated; Woodward, 1998; Onishi & Baillargeon, 2005). But for many years, our field lacked a quantitative framework that specified precisely how children and adults expect mental states to

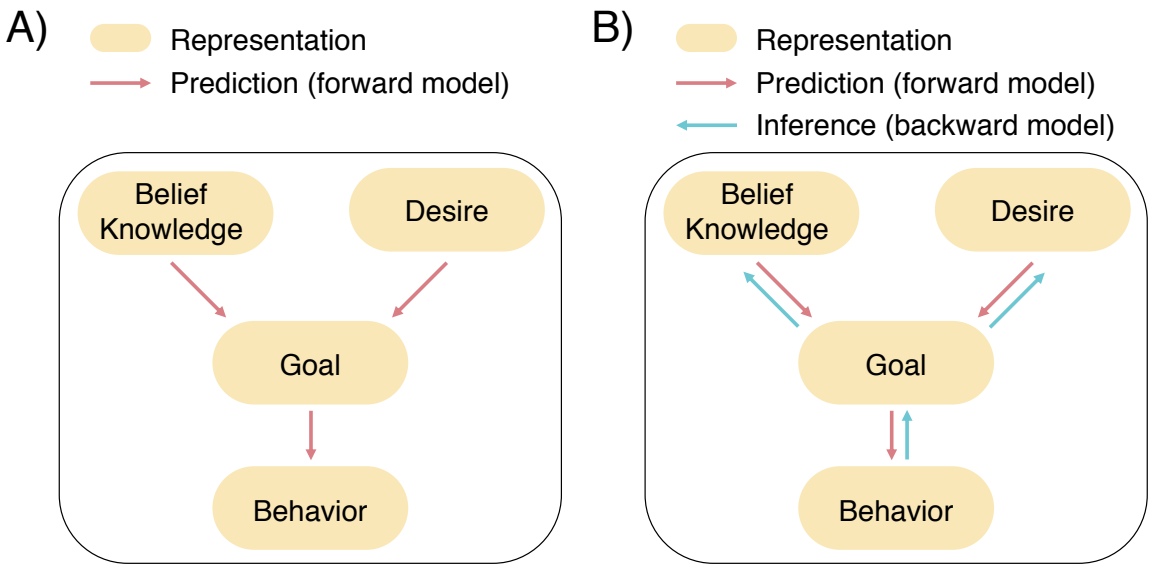

Figure 1.1: The basic components underlying mental state reasoning. A) Classic models of Theory of Mind focus primarily on children's mental state representations, and the feed-forward action predictions they enable; adapted from Wellman, 1992. B) More recently, the field has begun to focus also on mental state inferences from behavior, investigating how children and adults invert their causal models of other minds to infer what agents think from what they do.

lead to behavior (as represented by the pink arrows in Figure 1.1a). More recently, researchers have begun to focus more on this connective tissue, formalizing the mechanisms by which we use mental state representations to make action predictions.

Specifically, recent work suggests that our mental state reasoning is guided by a "naïve utility calculus", whereby we expect others to select action plans that maximize rewards, while minimizing costs (Jara-Ettinger et al., 2016). This expectation allows us to make principled action predictions given information about others' mental states: for instance, if Maxi is looking for chocolate, we expect him to continue searching as long as the reward of the chocolate outweighs the cost of finding it. Thus, if Maxi really wants chocolate (e.g., places a high reward on obtaining it), we can predict he will continue searching for quite a while. And if Maxi doesn't want chocolate very much (e.g., places a very low reward on obtaining it), we can predict that if he doesn't find it right away, he probably won't waste too much time searching.

As adults, however, we are not limited to simply predicting how others will act. Upon seeing agents' actions, we can invert our model of other minds to infer what someone believed or desired; see Figure 1.1b. For instance, if Maxi looks in the cupboard for his chocolate, we can infer he believed it was there; if he continues searching high and low, we can infer he really wants it; and if the chocolate is actually in another room, we can infer Maxi doesn't know that. Because the naïve utility calculus enables us to formalize how individuals make feed-forward action predictions, inverting this model also enables us to capture the process of mental state inference.

Recent models capitalizing on this framework have succeeded in quantitatively capturing adults' inferences about others' goals (Ullman et al., 2009; Jern & Kemp, 2015), desires (Velez-Ginorio, Siegel, Tenenbaum, & Jara-Ettinger, 2017), and preferences (Jern, Lucas, & Kemp, 2017), to name a few. Furthermore, recent research suggests that the expectations formalized by the naïve utility calculus are shared by both infants and adults (S. Liu et al., 2017; Jara-Ettinger et al., 2016), and can also capture children's predictions and inferences about others' goals (S. Liu et al., 2017), preferences (Jara-Ettinger et al., 2015; Lucas et al., 2014) and desires (Jara-Ettinger, Floyd, Tenenbaum, & Schulz, 2017). Taken together, this research suggests that human goal- and desire-reasoning is unified under an early-emerging expectation that others maximize utilities.

What remains less clear, however, is whether children and adults make action predictions and inferences over others' epistemic states in the same way. Relatively little research has formally investigated how adults infer others' epistemic states; and existing work has focused on fairly constrained cases where there are only a few things an agent could believe (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) or know (Jern & Kemp, 2015, Experiment 3). Similarly, little prior research has investigated how children infer what others know (Jara-Ettinger et al., 2017), or whether preschoolers are even capable of inferring beliefs from behavior (Wu & Schulz, 2018). Although early work in the Theory of Mind literature was concerned with the distinction between prediction and inference (see Wellman, 2011), these questions were largely put aside in the epistemic domain as the field began to focus on the false belief task as a conservative measure of Theory of Mind.

Thus, it remains an open question precisely how even adults make action predictions and inferences in the epistemic domain (especially in situations where there are many things an agent could believe or know)—let alone how these capacities develop. And in answering these questions, we may be able to shed light on the open puzzles posed above: revealing both what expectations and computations are shared across belief- and desire-reasoning, and showing what aspects of epistemic reasoning might still require time to develop.

1.4 Epistemic inference: A theoretical and computational framework

In this thesis, I present an expanded theoretical account of epistemic inference, which can capture not only inferences about *what* someone knows, but also how much. This account builds upon the naïve utility calculus, testing whether the expectation that others will maximize utilities also underlies epistemic reasoning in adulthood and childhood.

To review: under the naïve utility calculus, our ability to infer others' mental states relies upon an expectation that others will act efficiently, choosing action plans

that produce the greatest rewards while incurring the fewest costs (see Jara-Ettinger et al., 2016; Jara-Ettinger, 2019). Formally, see Equation (1.1):

$$Utility(plan, outcome) = Reward(p, o) - Cost(p, o) \tag{1.1}$$

While this framework can be applied to explain how humans reason about a variety of mental states, it does not capture the unique ways in which epistemic states can affect expected utilities. I show two ways this framework can be modified to better capture epistemic reasoning specifically. First, in most prior work, the cost of an action plan is formalized as a measure of the physical effort required to achieve an outcome (for example, the distance traveled to obtain a desired outcome). But intuitively, the cost of our action plan may depend upon what we know. For instance, if you're trying to catch a train at Grand Central and know exactly how to navigate to your track, you can go straight there; if you're not sure, it may take you more time and effort to find it. Thus, I propose that the cost of achieving an outcome can be affected by what and how much we know, K. Second, our ability to obtain our desired outcome (and receive its reward) can also depend upon what we know. For instance, your ability to catch the train before it departs may depend on how well you know where you're supposed to go; if you spend too long searching, you might miss it. Thus, I propose that the reward should be scaled by the probability we actually obtain it, given our knowledge state: $p(success|K)$. Formally, see Equation (1.2):

$$Utility(p, o) = p(success \mid K) \times Reward(p, o) - Cost_K(p, o) \tag{1.2}$$

Finally, we can often choose to seek added knowledge, $K+$. This knowledge may come at a cost (like a 6-year PhD!), but it can also modify both the probability of achieving our goals, and the difficulty of doing so. For instance, if you're really not sure where to go in Grand Central, you might decide to climb the stairs to the upper level, using your new vantage point to figure out exactly where your track is. While climbing the stairs confers an added cost, it may enable you to more easily and effectively navigate to your track. Formally, see Equation (1.3):

$$Utility(p, o) = p(success \mid K+) \times Reward(p, o) - Cost_{K+}(p, o) - Cost(K+) \tag{1.3}$$

As we navigate the world, then, we face a decision. Should we incur a cost to seek added information, or act based upon what we already know (see Figure 1.2)? Intuitively, rational agents should only act to seek knowledge when it is less costly than it is helpful. To predict what agents ought to do more formally, we can consider the cost differential between equations 2 and 3 (taking into account how costly added information will be, and how much it then reduces the cost of achieving our goal). We can also consider how much added information increases our probability of success (how helpful it will be), and how rewarding we find the outcome in the first place. The following equation captures the intuition that rational agents should only seek

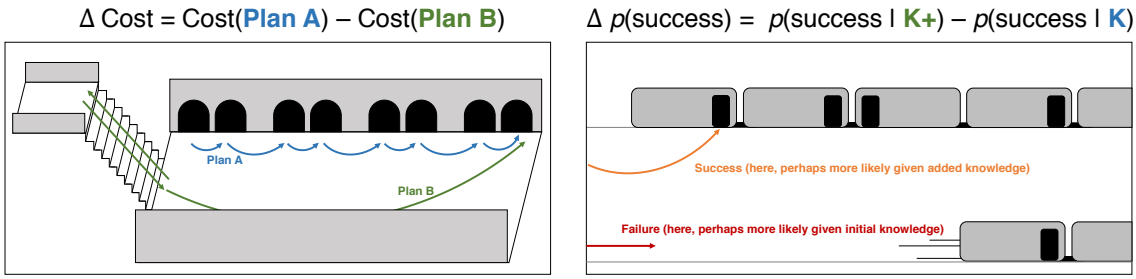

Figure 1.2: This figure breaks down two key terms from Equation (1.4). In the left panel, I contrast the cost of action plan A (searching Grand Central until finding our train) with the cost of action plan B (climbing the stairs to gain a bird's eye view, then going straight to our track). Given the number of tracks at Grand Central, it may be less costly overall to incur an initial cost to gain knowledge (climbing the stairs), rather than continuing to search without knowing where our track is. In the right panel, I depict two possible outcomes (successfully boarding the train; or failing, and having to wait until the next train arrives). What we know might affect the probability of each outcome. For instance, if we spend 10 minutes searching for our track, we might be more likely to miss the train than to catch it; and if we run up the stairs and manage to head straight to our track, our odds of success could be high. While the reward is not depicted here, this term simply expresses how much we care about the outcome we're pursuing.

knowledge when they expect the overall added costs to be exceeded by the reward. Formally, see Equation (1.4):

$$\Delta Cost < \Delta p(success \mid K) \times Reward(p, o) \tag{1.4}$$

By formalizing a theory of how people might make epistemic inferences, we can precisely predict how adults should respond if our theory is correct. If people respond as the model predicts across many situations, this suggests that participants use a similar process to make epistemic inferences. Thus, in this thesis, I use the above framework to understand both how adults make epistemic inferences, and how children's capacities develop.

1.5 Thesis overview

1.5.1 Chapter 2

In Chapter 2, I validated the above framework with adults, formalizing a model based upon these equations (e.g., formalizing our theoretical proposal), and showing that this model captures adults' epistemic inferences with quantitative precision. In a first experiment, participants were introduced to an agent going on Easter egg hunts (see Figure 1.3a). They learned that the agent sometimes had some knowledge over a field (for instance, they might know exactly where the prize is hidden, they may know several locations it is not hidden, or they may know nothing). In each trial,

the agent chose to search one of the two fields for the prize. Importantly, one egg in each field always contained an equivalent prize—so the agent would obtain the same reward no matter which field they searched. However, the expected cost of finding the prize could differ across fields: for instance, in Figure 1.3a, the eggs in Field A are all quite far from the entrance, and spread out. In Field B, the eggs are clustered together, and quite near the entrance. Thus, all else equal, it should be easier to search Field B.

After observing the agent's choice (and its expected cost), participants were asked to judge how much the agent knew about the location of the prize in each field. If adults expect others' costs to be modulated by their epistemic state, others' actions (and their expected costs) should sometimes reveal what they know. For instance, if an agent chooses to search the more apparently-costly Field A, the best explanation for this choice is probably that they knew precisely where the prize was in this field—and knew nothing about the prize's location in Field B (thus rendering Field A less costly to search than Field B overall). I found that our model well-captured adults' inferences, and did so reliably better than an alternate model which relied on a simpler heuristic, predicting the agent should always search the field they knew more about.

In a second experiment, we tested whether our model could capture participants' inferences not just about what agents know, but about what they believed they could discover. In this experiment, participants observed agents decide whether to search an island for treasure—or whether to obtain a treasure map first. Importantly, islands varied in size (some were small and thus easy to search even under ignorance; some

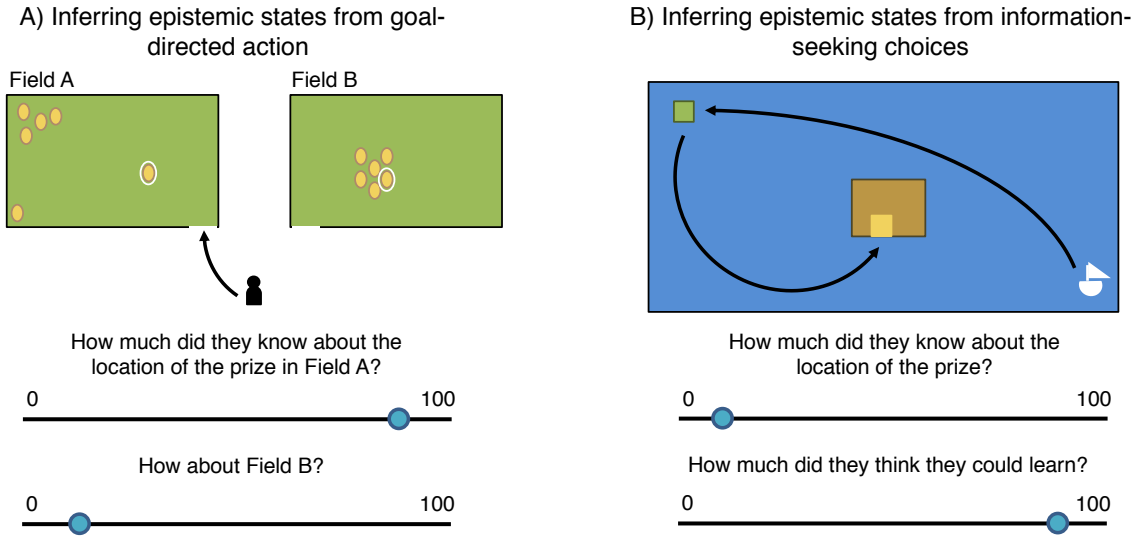

Figure 1.3: A) The arrow indicates the agent's chosen field; eggs containing a prize are circled. Here, participants might make a strong epistemic inference: the agent likely knows more about Field A (otherwise, they should've chosen to search the easier field). B) The arrow indicates the agents' chosen path. Here, participants might again make the strong inference that these agents don't know much, and believe they will learn a lot from the map.

were large), and the treasure map did not provide complete information (sometimes it held some information about the location of the prize, sometimes it held a lot, and sometimes it held none). After observing agents' information-seeking decisions (and the cost of these choices), participants were asked to infer both how much the agents already knew, and how much information they expected to gain from the treasure map. If adults expect agents to trade-off information's cost and reward, observing the costs agents incur (or reject) to gain knowledge should sometimes reveal what they know and believe they will learn. For instance, in Figure 1.3b, agents chose to take a costly detour to obtain a treasure map before searching a small island. From this, participants should infer both that the agents did not know very much about the location of the treasure, and expected to gain a lot of information from the map (otherwise, it would not have been worth the effort). Again, I found that our model well-captured adults' inferences.

1.5.2 Chapter 3

Given that our theoretical account captures adults' epistemic inferences, in the latter part of my thesis I use this account to investigate how these capacities develop. However, much prior research suggests that children decide whom to trust and believe via simple cues (like accuracy, familiarity, or confidence; see Mills, 2013; Kominsky et al., 2016). Thus, before testing how children make epistemic inferences, I test *whether* they leverage their Theory of Mind to make epistemic inferences in the first place (with the alternative being that young children instead rely on simple heuristics, rather than full mental state reasoning).

To test whether four- and five-year-olds make epistemic inferences in the absence of simple cues, participants watched two puppets tell them was inside two containers. One puppet looked, and then said what was inside each container. The other first said what was inside, and then looked. Participants were told that one puppet had

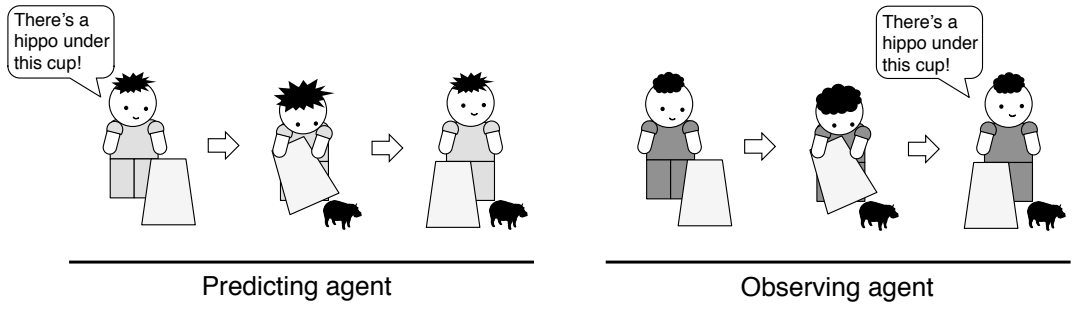

Figure 1.4: Procedure schematic for experiments presented in Chapter 3. Both agents perform identical actions. One simply says what's under the cup before he looks; the other first looks and then says what's there.

actually already peeked, and thus knew what was inside the containers. Although both puppets performed the exact same actions (just in the opposite order), by age five, children understood that the puppet who said what was inside the containers before he looked was probably the knowledgeable one. This suggests that young children can indeed make epistemic inferences in situations where simple cues will not suffice—opening the door to investigate how.

1.5.3 Chapter 4

In Chapter 4, I test one of the most basic predictions of our account. Prior research demonstrates that children expect others to maximize their utilities when pursuing external goals (see Jara-Ettinger et al., 2016, 2017). Do preschoolers extend the same expectation to the epistemic domain, expecting others to trade-off the costs and rewards of knowledge? To investigate, we introduced four- and five-year-olds to two puppets, and each was given a chance to lift a box to find out what toy was underneath. This action was difficult for one puppet (because she was weak) and easy for the other (because she was strong). In a first experiment, both puppets refused to lift the box, and participants were told that one puppet had already observed the experimenter placing the toy under the box. Participants judged that the strong agent had already known what was under the box—if she hadn't known, she should have just incurred a minimal cost to gain information. In a second experiment, both puppets agreed to lift the box, and participants were asked to judge which agent had the stronger epistemic desire. Here, participants judged that the weak agent had the stronger epistemic desire—because her choice to pursue knowledge came at a higher cost. These results suggest that by age four, children expect others to maximize their epistemic utilities; and thus, that by four years, mental state reasoning is already unified under an expectation that others will maximize their utilities.

1.5.4 Chapter 5

Given that even young children expect others to trade-off epistemic costs and rewards, the only components of our account that remain unexplored are the terms expressing an agent's probability of success given their knowledge state. In Chapter 5, I test whether preschoolers are able to estimate how likely an agent is to succeed on a task, given different degrees of knowledge. In a first experiment, four- to six-year-olds are introduced to an agent, and to eight boxes (four blue boxes, and four green boxes). Participants learn that on one side, every box has a marble underneath; and on the other side, only one box has a marble underneath. Participants are asked to gauge the agent's knowledge state by asking her to find a marble on one side. By age six (but not before), children preferred to ask about the side with a lower probability of random success, where there was only one marble. In a second experiment, participants were introduced to the same boxes, and to two agents. Each agent was asked to find a marble on one side. Both succeeded on the first try, and participants were asked

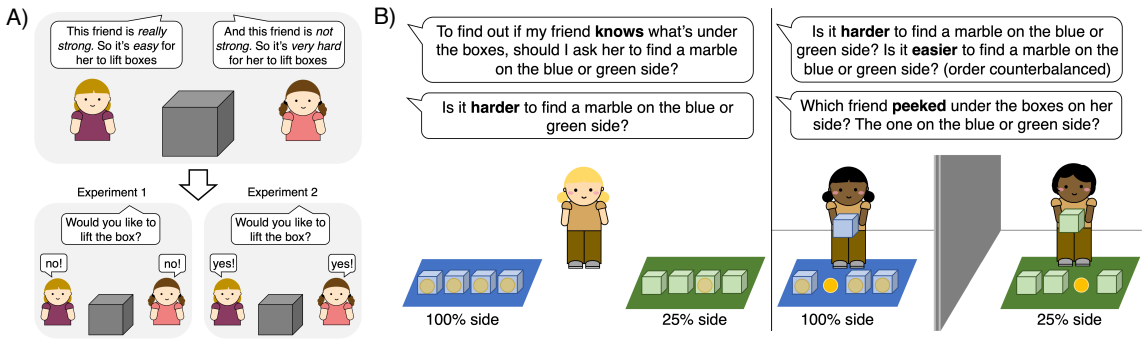

Figure 1.5: A) Procedure schematic for Chapter 4. In the first experiment, both agents refused to lift the box, and participants were asked who had already seen the experimenter hide the toy under the box; in the second experiment both agreed, and participants were asked who had the stronger epistemic desire. B) Procedure schematic for Chapter 5. In a first experiment, participants were asked to evaluate the agent's knowledge state by assigning her a task; in a second experiment, they were asked to judge which agent likely already knew what was in the boxes.

to judge which one had already peeked under the boxes, and knew what was there. Again, by age six (but not before) participants judged that the agent who found the only marble on her side was more likely to be knowledgeable (as compared to the agent on the other side, who would've succeeded no matter what. This suggests that by age six, but not before, children understand that when the probability of random success is low, only a knowledgeable agent is likely to succeed—but that when success is assured, the same outcome cannot diagnose an agent's epistemic state.

1.5.5 Chapter 6

Finally, I test whether children estimate how much added knowledge might increase an agent's probability of success on a task. In a first experiment, four- to six-year-olds observed two agents learn how a toy worked. Participants then saw each agent activate a new toy (outwardly identical to the first). One agent tried the same thing they learned made the first toy work; the other agent did something different. Both succeeded in activating their respective toy. By age five (but not before), children inferred that the agent who tried something different (and succeeded) must have already known how all the toys worked. A second experiment was similar to the first, except that now both agents failed. By age six (but not before) children inferred that the agent who rejected his experience with the first toy (trying something different) must have known more about the toys. These results suggest that by age six robustly (and by age five less reliably), children understand that ignorant agents will transfer relevant prior knowledge to novel situations—realizing that only an agent who has added knowledge is likely to succeed when faced with a situation where their prior experience unexpectedly does not apply.

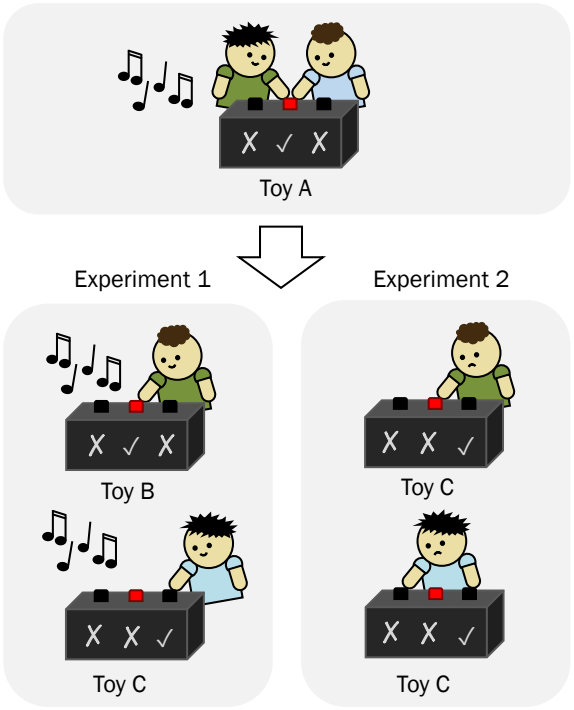

Figure 1.6: A) Procedure schematic for Chapter 5. Two agents learned how Toy A worked, and each was asked to activate a new toy. One agent always relied on his experience with Toy A when deciding what to do. The other rejected this experience, choosing to press a different button.

1.5.6 Chapter 7

Taken together, this work presents a theory-driven approach to understanding the core components of a Theory of Mind, and the computations that enable us to reason about others' epistemic states. In Chapter 7, I close with a discussion of my findings, and their implications for understanding not just epistemic reasoning, but also mental state reasoning more broadly.

Chapter 2

Validating a novel theoretical framework for epistemic inference

This chapter is based upon Aboody, Davis, Dunham & Jara-Ettinger (2021). I can tell you know a lot, although I'm not sure what: Modeling broad epistemic inference from minimal action. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society.*

Abstract

Inferences about other people's knowledge and beliefs are central to social interaction. In many situations, however, it's not possible to be sure what other people know because their behavior is consistent with a range of potential epistemic states. Nonetheless, this behavior can give us coarse intuitions about how much someone might know, even if we cannot pinpoint the exact nature of this knowledge. We present a computational model of this kind of broad epistemic-state inference, centered on the expectation that agents maximize epistemic utilities. We evaluate our model in a graded inference task where people had to infer how much an agent knew based on the actions they chose (Experiment 1), and how much they believed they could learn (Experiment 2). Critically, the agent's behavior was always underdetermined, but nonetheless contained information about how much knowledge they possessed or believed they could gain. Our model captures nuanced patterns in participant judgments, revealing a quantitative capacity to infer amorphous knowledge from minimal behavioral evidence.

2.1 Introduction

Imagine going to your friend's house for dinner and, as you're cooking together, realizing that you'll need more flour. As the two of you head out, you notice that your friend immediately starts walking in the direction of a large supermarket, rather than her usual go-to bodega around the corner. From this simple decision you might quickly suspect that she knows something you don't. Perhaps the bodega doesn't carry flour; maybe it's cash only and your friend intends to use her credit card; or the supermarket might be the only place that's open late. Inferences like these not only enable us to make sense of others' behavior, but also help us decide when to share what we know, and from whom to learn what we don't, forming a cornerstone of complex social action.

The ability to interpret other people's behavior in terms of mental states, called a *Theory of Mind*, has its origins in early childhood. From infancy, we interpret other people's behavior as goal-directed (Woodward, 1998) and infer others' goals and preferences by assuming that agents act to maximize utilities—the difference between the costs they incur and the rewards they obtain (Csibra, 2003; Jara-Ettinger et al., 2016; S. Liu et al., 2017). Throughout our life, this expectation enables us to make a variety of judgments, such as inferring what others like (Lucas et al., 2014; Jern et al., 2017), predicting how they might behave (Jara-Ettinger, Schulz, & Tenenbaum, 2020), and determining their social affiliations (Jern & Kemp, 2014; Ullman et al., 2009; Jara-Ettinger et al., 2015).

As the example above shows, however, inferences about others' minds are not restricted to goals and preferences: they also include judgments about what others may or may not know. Consistent with this, research in computational social cognition has found that people can make quantitative inferences about others' beliefs based on their behavior (Baker et al., 2017). This work showed that a computational model of joint belief-desire attribution, embedded in a Bayesian framework for action understanding, captures how people determine what an agent believes about their environment given their behavior (e.g., if an agent looking for lunch walks towards the end of the block, peeks around the corner to see a Mexican food truck, and then turns around, we can infer that the agent was hoping to see a different food truck there).

While this work shows that people can make targeted belief inferences, such as determining whether an agent knew the type of food a vendor might be selling based on their behavior, these inferences often require access to a relatively constrained hypothesis space and key actions that reveal the agent's beliefs. In many everyday situations, however, we may not immediately know the most relevant epistemic hypothesis to consider, and other people's behavior may not contain the amount of information needed to disambiguate between different degrees of knowledge. In cases like these, our representations of other people's epistemic states appear to consist of amorphous estimates of how much others know, without being sure exactly what it is that they know. Returning to the example in the introduction, when your friend

chose to go to the supermarket it was easy to infer that she knew more than you did, despite not knowing exactly what she knew. What computations underlie such epistemic inferences?

Research investigating the ability to estimate and quantify how much others know—intuitions about how much people know without specifying exactly what they know—has generally focused on children. By early in preschool children can represent how much others know about a domain, without needing to list the full contents of their knowledge (Landrum & Mills, 2015; Lutz & Keil, 2002). However, to our knowledge, no work has explored our capacity to infer knowledge magnitude from others' actions, or specified the computations that might underlie such inferences.

Here we propose that such inferences are supported by an expectation that agents maximize utilities, through a sensitivity to the apparent costs that agents choose to incur. Specifically, we suggest adults understand that the costs agents incur often depend on the knowledge they possess. Thus, an ability to infer the subjective costs that an agent appears to act under can reveal the amount of knowledge they might have. In the example above, for instance, the fact that your friend chose to incur a seemingly higher cost (walking to a place that was farther away) for the same reward (getting flour), suggests that she possessed privileged information—leading her to conclude that the large supermarket was a better option than you'd originally assumed.

In this paper we present a computational model of epistemic quantification through an expectation that agents maximize utilities, and we test its performance on tasks where participants must infer how much someone knows or thinks they can learn based on their behavior. Our work shows that people can seamlessly make graded quantitative estimates of how much someone knows or expects to learn, and that these inferences can be explained through an expectation that agents maximize utilities (the difference between the costs they incur and the rewards they obtain), and an understanding that the costs agents incur depend on the knowledge they possess.

2.2 Computational Framework

Our computational framework builds on a recent family of computational models of mental-state inference structured around an expectation that agents act rationally—formalized as a generative model of utility maximization, combined with a mechanism for inverting this causal model via Bayesian inference (Lucas et al., 2014; Jern et al., 2017; Baker et al., 2017; Jara-Ettinger et al., 2020). We extend this framework by proposing that adults often expect agents' costs to be mediated by their knowledge state—and can thus infer others' epistemic states from observing the apparent costs they choose to incur.

In this project, we consider scenarios in which an agent's knowledge state affects the expected cost of an action plan. For example, consider the scenario from the introduction: if an agent is 50% certain that the (closer) bodega has flour, and 100%

certain that the (further) supermarket has flour, then the expected cost of the action plan "go to bodega, then if no flour go to supermarket" is [distance from home to bodega]+ .5*[distance from bodega to supermarket], while the expected cost of going to the supermarket first is just [distance from home to supermarket]. The optimal action in this case therefore depends on both a) how much further away the supermarket is than the bodega and b) the agent's certainty that the bodega has flour.

In general, for a given action plan a, we can write $E[Cost(a)|k]$ to mean "the expected cost of implementing action plan a, given knowledge state k". If the reward of completing action plan a is R(a), then the total expected utility of action plan a, given knowledge state k, is $R(a) - E[Cost(a)|k]$.

Suppose now that the agent has knowledge state k and set of possible action plans $\{a_1, \ldots, a_n\}$. A standard assumption for utility-based agent models is that the agent will compute the expected utility of each action plan $R(a_i) - E[Cost(a_i)|k]$, and probabilistically select the best option through softmaxing. By placing a prior distribution $P(k)$ over possible knowledge states $k \in K$, a Bayesian observer can make inferences about the agent's knowledge state k, given the observed actions a according to Bayes' rule; see Equation (2.1) below.

$$P(k|a) \propto P(a|k)P(k) \tag{2.1}$$

This model allows observers to infer a posterior distribution over what an agent knows given their behavior. As the example in the introduction shows, however, many situations are under-determined, and in these cases it is not possible to infer exactly what an agent knows. In our framework, this situation arises when an agent's behavior is consistent with a range of different knowledge states, and it is thus impossible to determine under which exact knowledge state the agent was acting. But even when we can't infer the precise contents of others' knowledge representations, we may still be able to infer approximately how much they know (getting a rough sense of how knowledgeable they are). Thus, given a posterior distribution over what the agent might know, we formalize the quantity of amorphous knowledge Q as the expected quantity of knowledge encoded in the probable epistemic states that the agent has, given by Equation (2.2) below.

$$Q = \sum_{k \in K} |k| p(k|a) \tag{2.2}$$

where K is the set of all possible epistemic states, $|k|$ is a quantification of how much the agent knows in that state, and $p(k|a)$ is the posterior probability of that knowledge state (Eq. 2.1). Naturally, precisely defining the measure $|k|$ may be highly context-sensitive. Here we focus on its application in a particular experimental context but return to the idea of how this might generalize in the discussion.

To evaluate this framework, we considered cases like those shown in Figure 2.1. Here, an agent is deciding which of two fields to go on an easter egg hunt in, knowing that each field has only one egg with a prize inside (and that the reward is always

the same in every field). In this paradigm, a knowledge state k consists of a subset of eggs that the agent might be aware of. The cost the agent incurs depends on their search trajectory. We assume that agents navigate efficiently in space (Csibra, 2003), and search only locations where they think they may find the prize (that is, agents should not visit a location they are sure does not contain the prize). This implies that when the agent's knowledge state includes information about the contents of the prize egg, the agent will always move directly towards the prize. Otherwise, the agent will search the eggs in a way that minimizes the expected search time. Finally, to compute equation 2.2, we treat the amount of knowledge in an epistemic state as 1 − the proportion of eggs (or island squares) the agent is still uncertain about (if the agent knows where the prize is, they know the rest of the eggs or island squares are empty, and thus the proportion known is 1; if the agent is unsure about half of the eggs or island squares, the proportion known is .5; and so on).

To produce a prediction for a given scenario in Experiment 1, our generative model sampled 10,000 knowledge states over the contents of each field, predicted the agent's actions by softmaxing the expected utility of each action plan given the agent's knowledge, and recorded both the agent's knowledge state, and the agent's predicted choice. To perform epistemic inference, when given an agent's actual choice, the model returned the average proportion agents knew under cases where the predicted choice was consistent with the observed choice. In Experiment 2, rather than sampling knowledge states, we instead enumerated every possible set of epistemic states (how much agents themselves knew, and how much they believed they could learn). We predicted agents' actions by softmaxing the expected utility of each action plan given the agent's knowledge and beliefs over information gain. We perform epistemic inference by obtaining the expected knowledge and belief state through marginalization (we also defined priors over how much agents generally know, and how much information maps usually contain; see Experiment 2 Model Parameters).

2.3 Experiment 1

2.3.1 Model Parameters

Our main model has four parameters: the reward of obtaining the prize, the cost of checking an egg's contents upon reaching it, a prior over the agent's knowledge, and the softmax parameter (τ). All parameter values and model predictions were preregistered prior to data collection (link to OSF preregistration: `https://osf.io/95qxk/`).

The reward function for the prize is the same across fields, and we set it as a constant $R(a_i) = 100$. Because the reward is constant across action plans, the difference in utilities between the two plans would be unchanged by different reward functions. We simply selected (and preregistered) a reward function large enough to ensure that no action plan could have a negative utility.

For each knowledge state sample, the cost of stopping to check an egg's contents was randomly chosen from the continuous uniform distribution $[1, 3]$. This range was chosen to capture the expectation that stopping to open an egg does incur some cost, but that this cost is relatively minimal.

We specified a prior over the agent's knowledge: the agent had a 50% chance of knowing each egg's contents. We also explicitly communicated this to participants in our task (see Procedure) to ensure that participants and the model both relied on similar epistemic priors. Finally, we selected a softmax τ value that produced graded action predictions in proportion to each plan's expected utility ($\tau = 3$).

Our alternate model was not preregistered, but uses only one parameter: the same knowledge prior as in our main model. Because our alternate model encodes an expectation that agents will always choose fields they know more pieces of information about, we do not compute the utility of each field, and thus we do not need to specify agents' costs, rewards, or a softmax parameter.

2.3.2 Alternate Model

Our main model assumes that people quantify the cost of obtaining the prize in each field under different degrees of knowledge, and then reason about the knowledge states under which the agent's actions would have been utility-maximizing. However, it is possible that adults generally do not apply such complex computations when inferring others' knowledge states, and instead rely on simpler rules or heuristics. Such heuristics could get things right most of the time, while requiring less effort to apply.

To address this possibility, our alternate model encoded the simple heuristic that agents tend to choose options they know more pieces of information about. Critically, this alternate model did not consider agents' knowledge states in a full mentalistic way: it did not compute the utility of each field based on the agent's knowledge state, and did not expect agents to navigate directly to an egg if they knew it contained the prize. It simply considered the proportion of eggs with known contents in each field, and expected the agent to always choose the field where this proportion was larger (or choose randomly when this proportion was equal across fields). We then generate predictions from this alternate model using the same sampling procedure as in the main model.

2.3.3 Experiment 1 Rationale

To test our model, we designed a task where an agent's behavior (and its costs) could reveal approximately how much they knew—but was too impoverished to reveal precisely what they knew. Specifically, participants watched an agent choose which of two fields to search for a prize hidden in an easter egg. The cost of locating the prize in any given field was determined by the number of eggs, their spatial distribution, and the true location of the prize. By manipulating all three variables, we test if

participants infer how much others know by quantifying and comparing their expected costs—or whether participants rely on a simpler heuristic that does not require them to track or reason about others' costs when inferring epistemic states. Our procedure, stimuli, sample size, and analysis plan for our main model were preregistered (see OSF preregistration).

2.3.4 Participants

40 adult participants with U.S.-based IP addresses were recruited via Amazon Mechanical Turk ($M = 35.05$ years, $SD = 9.23$). 7 additional participants were recruited but excluded from the study for failing a preregistered inclusion trial.

2.3.5 Stimuli

Stimuli consisted of 19 test trials, plus one inclusion trial. The test trials were presented in a randomized order, and the inclusion trial was always presented last. Each trial showed an agent, and two fields. The fields each had easter eggs placed inside, and one egg in each field contained a hidden prize. This egg was circled for participants. An arrow indicated the agent's path to their chosen field, thus showing which field the agent chose to visit on each trial (see Figure 2.1).

Stimuli were based on three scenarios (pairs of fields) we thought could elicit a range of model ratings. To manipulate the cost of searching each field, eggs in the first field (field A) were always wide-spread. The second field (field B) contained the same number of eggs, but these eggs were instead clustered near the middle of the field. The first scenario is shown in Figure 2.1a. The second scenario was based on the first: we selected a subset of 6 eggs from each field, thus varying the number of eggs but not their position. The third scenario was in turn based on the second, but here we instead varied the position of the eggs in field A (capturing a case where most of the eggs in field A were extremely costly; see Figure 2.1b).

To select the final locations of the prize in field A, we provided each scenario as input to the model, but systematically varied which egg in field A contained the prize, yielding 42 trials (21 unique scenarios x 2 choices per scenario).[1] We selected 24 trials (12 unique scenarios x 2 choices per scenario) that both produced a range of model responses, and were not too similar to each other. In preparation to present stimuli to participants, some trials were mirrored, and we slightly varied the position of the prize in field B amongst similar scenarios (to prevent participants from noticing similarities between trials).[2] We then obtained final model predictions, and excluded any trials

[1]We did not expect the location of the prize in field B to strongly affect the model's predictions; to test if this was the case, we did also replicate one scenario given a different prize location in field B, yielding an additional 18 additional trials. The location of the prize in field B indeed had little effect (as all of these eggs are so close to each other), and thus we selected our final stimuli by considering primarily the location of the prize in field A.

[2]Despite slightly varying the prize's location in field B across similar trials in our preregistered stimuli, our model predictions were not updated accordingly prior to preregistration. Because we

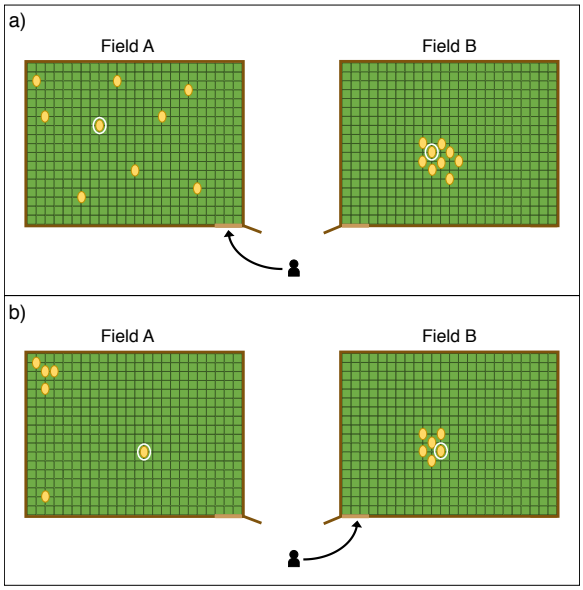

Figure 2.1: Example of the experimental stimuli. The arrow indicates the agent's chosen field; eggs containing a prize are circled. Panel A depicts a strong epistemic contrast: here, you might infer that the agent knows approximately where the prize is located in their chosen field, and very little about the other field. Panel B depicts a more graded contrast: here, you might suspect that the agent knows more about the prize's location in their chosen field, but may be less certain they know a lot (because their chosen field is also much less costly to search).

where the model's knowledge predictions were based on less than 500 samples (that is, where the predicted choice of field was consistent with the observed choice in less than 5% of cases). This yielded 19 final trials; this criterion and our final set of stimuli was preregistered.

2.3.6 Procedure

Participants were introduced to an agent going on easter-egg hunts in a two-dimensional grid-world. Participants learned that a farmer had placed easter eggs in his fields, hiding a prize inside one egg in every field. This prize (one silver token) was always the same in every field, and the prize egg was always circled for participants.

Participants learned that because the grass in the fields was quite short, the agent could always see where the eggs were located in a field before entering it. But while the prize egg was circled for participants, the agent didn't necessarily know which egg contained the prize. Participants learned that the agent had seen the farmer set up some of the eggs; it was unclear what prior over knowledge participants would bring to the task, so we specified that the agent had a 50/50 chance of knowing the

collected our data using the preregistered stimuli, we obtained new model predictions for any trials where the location of the prize in the stimuli did not match the coordinates originally used to generate the preregistered model predictions. We used the same preregistered parameters.

contents of any given egg. And participants were explicitly instructed that the agent did not always know the same amount about every field; the amount she knew about the location of the prize in each field could differ.

Participants learned that the agent always had to choose between two fields, and could only search the field she chose. An arrow indicated which field the agent had chosen to search (see Figure 2.1). Participants were oriented to factors that might affect the agent's search decision: they were told that the agent always wanted to find the prize as quickly and easily as possible, and that the difficulty of finding the prize was determined by the number of eggs in a field, their distance from the entrance, and the amount the agent already knew about the location of the prize. Note that while this tutorial ensured participants were attentive to the main features of our task, we are interested in how participants combine these different pieces of information and reason over them to infer what others know. The tutorial did not specify how participants should weight or use any of these features in their judgments.

To access the task, participants then completed a preregistered inclusion quiz that assessed their understanding of the task instructions. Participants were given two chances to pass the inclusion quiz; those who failed on their first attempt were required to review the task introduction before trying again. Participants who failed both attempts were not given access to the task. Upon passing the inclusion quiz, participants then completed the 19 test trials (presented in a randomized order), plus one inclusion trial at the end. For each trial, participants were asked to rate, on a sliding scale from 0 - 100, how much the agent knew about the location of the prize in each field. Critically, participants rated how much the agent knew about both fields, not just the field she had chosen. The preregistered inclusion trial always came last. It was similar to the test trials, but presented an extreme contrast where we could make a strong prediction about the pattern of judgments an attentive participant should make. Participants whose judgments differed from our preregistered criteria were excluded. Finally, participants were asked what they thought the point of the task had been, and were given an opportunity to provide feedback or note any technical difficulties.

2.3.7 Results

Participants rated the agent's knowledge about both fields in 19 test trials, yielding 38 ratings. As preregistered, participant responses were averaged by question, and then z-scored; the corresponding model predictions were also z-scored.

Figure 2.2 shows the overall results, revealing that our model was highly correlated with participant judgments, $r = 0.94$ (95% CI: $91.78, 98.84$). And this correlation did not reflect only cases where both the model and participants inferred a lot of knowledge or very little knowledge. Critically, it included cases where both the model and participants were equally uncertain, in a graded manner, about how much the agent knew. Figure 2.3 plots the trial-by-trial correspondence between model and participant ratings, showing that participants' judgments were not bi-modal, but

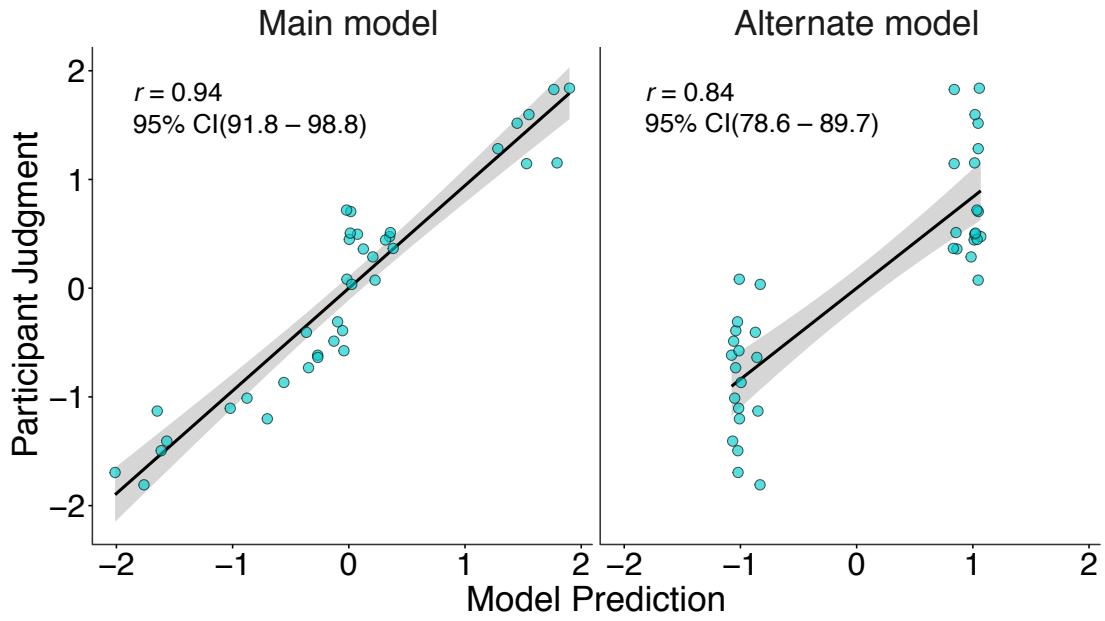

Figure 2.2: Comparison between our model and the alternate model, with linear regressions fit to each dataset. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression.

rather graded in a way that closely tracked our model's predictions.

To ensure that these results could not be the product of a simple heuristic, we implemented an alternate model. Rather than performing full mental-state inference, our alternate model simply assumed that agents always choose fields where they know about a greater proportion of eggs. Note that we only preregistered an analysis plan for our main model, but test the performance of the alternate model in the same way. The alternate model showed a weaker correlation with participant judgments, $r = 0.84$ (95% CI: $78.58, 89.77$), demonstrating that the amount of locations an agent knows about in each field does matter, but that predictions made on the basis of this one factor (without considering costs) do not capture the full pattern of participant judgments. A bootstrap over the correlation difference revealed that the main model was reliably better correlated with participants judgments than the alternate model (correlation difference, alternate model − main model $= -0.11$, 95% CI: $-17.34, -4.36$; not preregistered). As Figure 2.2 reveals, although the correlation between the alternate model and participant judgments was still high, this is only because the alternate model categorized every judgment into two rough bins. These predictions were approximately correct, but lack the nuance that participants' epistemic inferences showed, and that our model was able to capture.

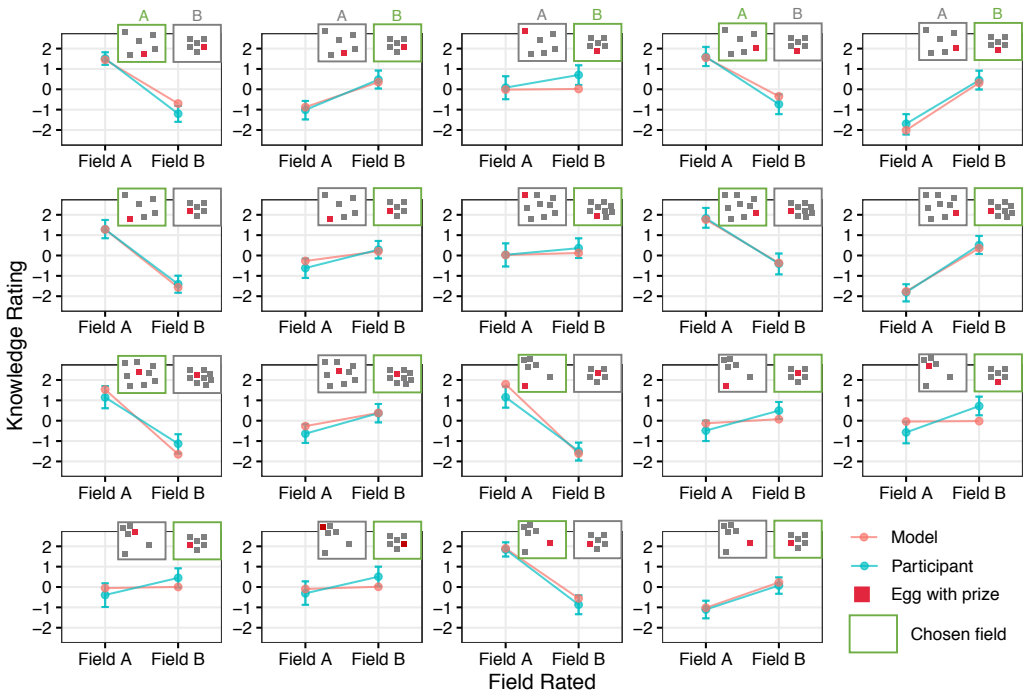

Figure 2.3: Detailed results for the experiment. Each panel presents one trial, with results split by the field rated (Field A or Field B, indicated on the x axis). The y axis indicates standardized knowledge ratings. Participant judgments are plotted in blue; model predictions are plotted in red. Vertical bars show 95% confidence intervals over participant judgments. The schematics show the position and number of eggs in each field, the egg with the prize, and the field the agent ultimately chose in each trial.

2.4 Experiment 2

2.4.1 Model Parameters

Our main model for Experiment 2 has six parameters: the reward of obtaining the prize (set to a constant $R(a_i) = 100$), the cost of sailing across one grid square, the cost of searching one island square, the softmax parameter (τ), and two priors: one over how much agents know in general, and one over how much information maps generally hold.

We pre-registered the first four parameters prior to data collection, basing the relative cost of sailing vs. searching upon empirical estimates from a pilot sample. Our pilot sample judged that searching one island square was, on average, 2.25x more difficult than sailing across one ocean square, and thus we pre-registered a sailing cost of 1, a searching cost of 2.25, and $\tau = 4$ (based upon the range of utilities these costs produced). However, we explicitly pre-registered that we would re-estimate these based on our final sample, and re-adjust our softmax parameter if needed. In our final sample, most participants judged that searching was more difficult than sailing, judging that it was on average 3.9x harder. Thus to generate our final predictions, we set the cost of searching to 3.9. Because this affected the range of possible utilities,

as preregistered we adjusted our softmax parameter, setting $\tau = 6.5$.

We also defined a uniform prior over the probability that the map might contain each degree of knowledge, and defined a non-uniform prior over the probability that the pirates might have each degree of knowledge (not preregistered). This was intended to capture the possibility that adults might generally expect agents to be knowledgeable (and unlike in Experiment 1, we did not specify precisely how likely agents were to know the contents of each island square). We defined this prior using the binomial distribution ($p = 0.8$).

2.4.2 Alternate Model

Our preregistered alternate model is a linear regression, trained on participants' z-scored average ratings in our task. It predicts knowledge based on an interaction between agents' information-seeking choice (to retrieve the map / skip the map), and the type of knowledge (what agents know / what information they believe the map contains). The formula for this regression in R is: `lm(mean participant rating ~ choice*knowledge category)`.

2.4.3 Experiment 2 Rationale

Experiment 1 shows that adults are able to make precise epistemic inferences even in underdetermined scenarios—and that these inferences are well-captured by our main model. Experiment 2 both conceptually replicates and extends these findings. Specifically, in Experiment 2 we test whether our framework can capture not just adults' inferences about how much someone knows, but also about how much they believed they could learn. To do so, we designed a task where an agent's information-seeking choice (and its cost) could reveal approximately how much they knew and believed they could learn (but again, could not reveal these states with any precision). Specifically, participants watched agents search islands for hidden treasure. Agents had the option to obtain a treasure map first, or to skip the map and go straight to the island. Importantly, the map was not always informative: sometimes it might contain a lot of information about the treasure's location, sometimes it might contain a little, and sometimes it might contain no information at all. To elicit graded inferences, we manipulated the distance of the map (varying information's cost), the size of the island (varying the potential difficulty of finding the treasure), and agents' information-seeking choices (varying whether or not they pursued the map). Our procedure and sample size were pre-registered.

2.4.4 Participants

40 adult participants with U.S.-based IP addresses were recruited via Amazon Mechanical Turk ($M = 38.73$ years, $SD = 12.23$). 9 additional participants were recruited but excluded from the study for failing a preregistered inclusion trial.

2.4.5 Stimuli

Stimuli consisted of 18 test trials, plus two inclusion trials. The test trials were presented in a randomized order, and the inclusion trials were always presented last. Each trial showed a pirate ship (represented by a yellow star), a treasure map (represented by a green square), and an island (represented by brown squares); see Figure 2.4. Each island had a beach (represented by a lighter brown square), which was the only point on the island pirates could land their ship. An arrow indicated agents' path, showing whether they chose to pursue added knowledge (obtaining the treasure map first), or whether they chose to search the island without obtaining the map (see Figure 2.4a).

To construct our stimuli space, we varied the size of the island pirates needed to search (12, 24, or 36 grid-squares), the detour required to obtain the treasure map (adding approximately 10, 20, or 40 grid-squares to the journey), and agents' choices to obtain or skip the map. This yielded 18 test trials which systematically varied information's cost (as well as agents' information-seeking choices).

2.4.6 Procedure

Participants were introduced to pirates searching for treasure in a two-dimensional grid-world. Participants were shown how to identify the pirate ship (marked by a star), and learned that pirates could only land on the island at the beach (this was intended to explain why the pirates sometimes took circuitous, high-cost paths to the island; e.g., see Figure 2.4a). Participants learned that pirates sometimes knew a lot about the treasure's location, sometimes knew a little, and often knew something in between.

Participants learned that islands could be all different sizes, and that there was always a map somewhere in the ocean, marked by a green square. However, this map was not always helpful: sometimes it contained a lot of information about the location of the treasure, sometimes it contained only a little, and often it contained something in between. To obtain the map, pirates needed to sail to the green square first, before going to the island. An arrow indicated pirates' final choice (showing their chosen path).

Participants were oriented to factors that might affect agents' information-seeking decisions: they were told that the less pirates knew, the more work it might take to locate the treasure; the bigger the island, the more work it might be to search for treasure; and the farther the map, the more time and effort might be required to obtain it. Participants were explicitly told that, in each case, the pirates needed to decide whether it was worthwhile to pursue the map. As before, note that while this tutorial ensured participants were attentive to the main features of our task, we are interested in how participants combine these different pieces of information and reason over them to infer what others know and believe they can learn. The tutorial did not specify how participants should weight or use any of these features in their

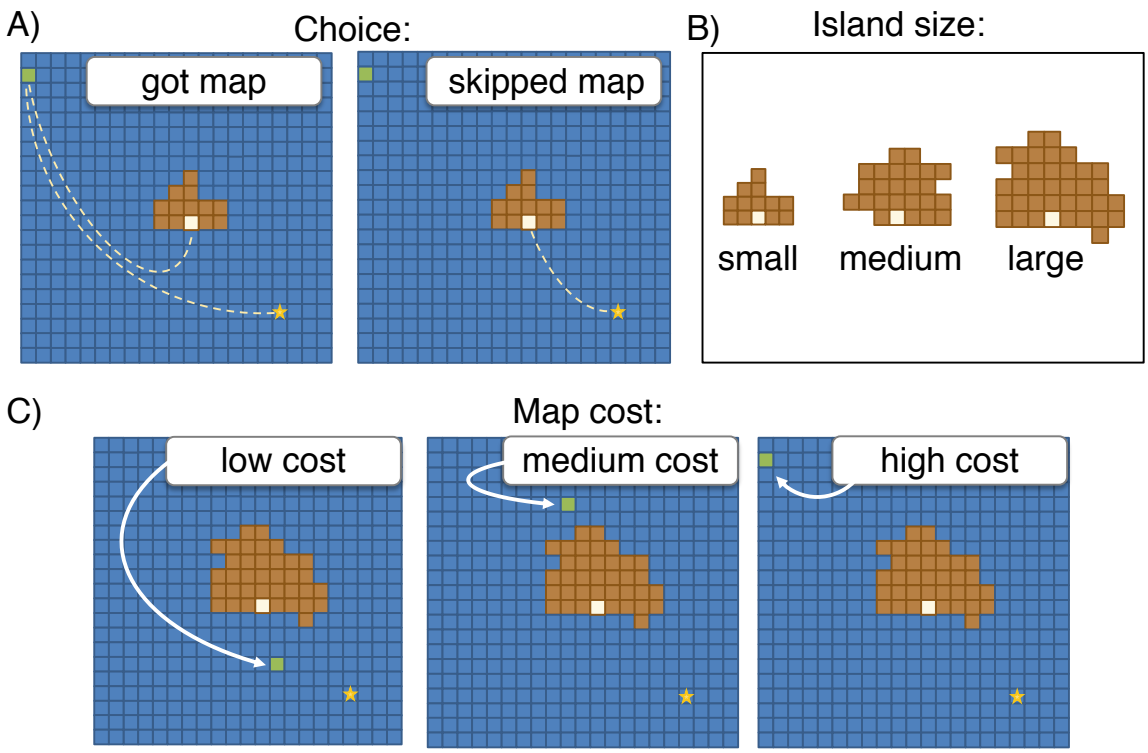

Figure 2.4: Space of all possible experimental stimuli. A) We varied agents' choices (to pursue information or ignore it), B) the size of the island to be searched (small, medium, large), and C) the cost of pursuing information (small, medium, large). This yielded 18 test trials. The first choice (panel A, left) depicts a strong epistemic contrast: here, you might infer the agents knew relatively little, and believed they stood to gain a lot of information (because they chose to incur a high cost to obtain the map, even though the island was small and thus relatively easy to search). The second choice (panel A, right) depicts a more graded contrast: while the agents clearly did not think the map was worth it, it may not be entirely clear why (did they know a lot, or did they simply believe the island would be easy to search even given ignorance?)

judgments.

Before the task, participants completed three simple attention check questions that assessed their understanding of the task instructions. Participants were asked to identify how the pirate ship was marked (by a star), to recall the pirates' goal (find treasure), and finally were asked to identify both that the map was always on the green square, and that pirates could only get on an island via the beach (distinguishing these from three other incorrect statements). Participants were able to select as many answers as they chose to each question; however, attentive participants should have noticed that the first two questions could only have one correct answer. Any participants who selected more than one answer in response to these two questions was excluded (preregistered). Participants who answered any question incorrectly were corrected.

Finally, participants were again reminded that both the pirates' knowledge and the

informativeness of the map might vary, and that in each case, pirates needed to decide whether it was worthwhile to pursue the map. For each trial, after observing pirates' information-seeking choices (and their expected costs), participants were asked to rate, on a sliding scale from 0 - 100, how much the pirates knew about the location of the treasure, and how much information the pirates thought the map had about the location of the treasure.

Two inclusion trials always came last. These were similar to the test trials, but presented an extreme contrast where we could make a strong prediction about the pattern of judgments an attentive participant should make. Participants whose judgments differed from this pattern were excluded, as preregistered.

Participants were also asked to judge which was more difficult: to sail across one ocean square, or search one island square for treasure. After identifying which was harder, participants were asked to judge how much more difficult their chosen option was, in relation to the other. This choice was preregistered, with the idea that the cost our model assigned to each action (sailing vs. searching) would be scaled based upon participants' judgments. Finally, participants were asked what they thought the point of the task had been, and were given an opportunity to provide feedback or note any technical difficulties.

2.4.7 Results

Participants rated how much the pirates knew, and how much they believed they could learn from the map, in 18 test trials. This yielded 36 final ratings. As in Experiment 1, participant responses were averaged by question, and then z-scored; the corresponding model predictions were also z-scored.

Figure 2.5 shows the overall results, revealing that our model was highly correlated with participant judgments, $r = 0.86$ (95% CI: $89.93, 92.7$). And this correlation did not reflect only cases where both the model and participants inferred a lot of knowledge or very little knowledge. Critically, it included cases where both the model and participants were equally uncertain, in a graded manner, about how much the agent knew.

To ensure that these results could not be the product of a simple heuristic, we implemented an alternate model. Rather than performing full mental-state inference, our alternate model simply assumed that an agent who skipped the map didn't need information, and vice versa. Because this model was insensitive to cost, it did not consider more graded cases we expected humans might (e.g., that if the map is right on the way you might check even if you're not sure how much you'll learn; whereas if the map is far away, you may choose not to obtain it even if you lack some knowledge). This alternate model showed a stronger correlation with participant judgments, $r = 0.94$ (95% CI: $91.3, 97.8$); a bootstrap over the correlation difference revealed that the alternate model was reliably better correlated with participants judgments than the main model (correlation difference, alternate model − main model = 0.079, 95% CI: $0.9, 14.6$; not preregistered).

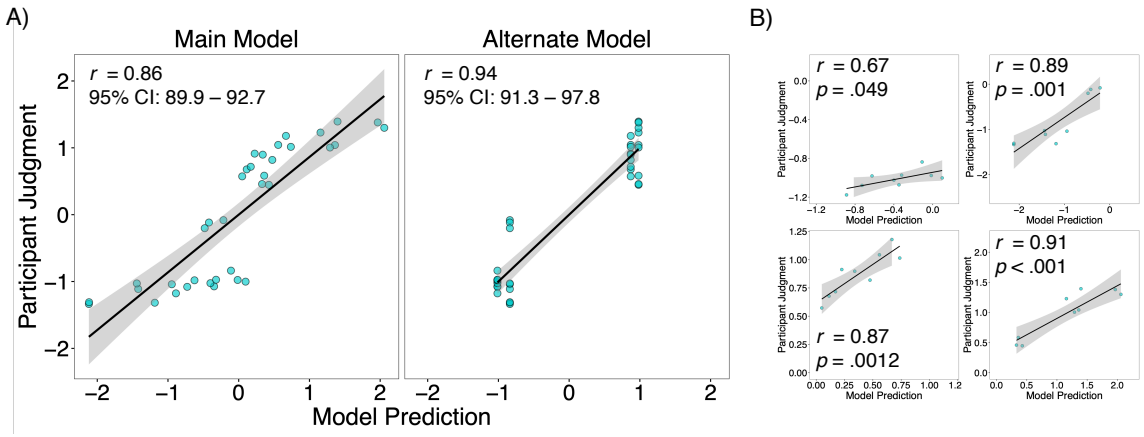

Figure 2.5: A) Comparison between our model and the alternate model, with linear regressions fit to each dataset. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression. B) Correlation between participant judgments and main model, binning participant judgments according to the alternate model's predictions. Each point represents one knowledge rating, with model predictions on the x axis and participant judgments on the y axis. Gray bands show 95% confidence intervals in the regression. This reveals meaningful variation our alternate model was not able to capture.

Although the alternate model was better correlated with participant judgments (perhaps not unexpectedly, as it was trained on participant judgments in the first place), it did not capture any of their gradedness. While it is generally true that in our task, agents sought out information when they needed it and skipped it when they did not, both participants and our main model were able to make much more nuanced epistemic inferences. Thus, following our preregistered analysis plan, we test whether there is actually meaningful variation in participant judgments that the alternate model fails to capture (despite well-capturing the overall trajectory of participants' responses).

Specifically, because the alternate model binned all predictions into four categories, we tested whether participant judgments *within* each of these categories still well-correlated with those of our main model. If this is the case, this would suggest that the alternate model fails to account for meaningful variation. In other words, obtaining meaningful correlations within each bin suggests that there is still structure in each category that only our main model is able to capture. Consistent with this possibility, even when separating participant judgments according to the predictions of our alternate model, participants' judgments were significantly correlated with the corresponding judgments from our main model (all r's between $[0.67, 0.91]$, all p's $< .05$; see Figure 5.5). This demonstrates that our alternate model fails to capture meaningful variation in participant judgments, despite the high overall correlation between participant judgments and the predictions of our alternate model.

2.5 General Discussion

Here we presented a computational model of amorphous epistemic inference. Our model sought to explain people's capacity to infer how much others might know or expect to learn in situations where inferring their precise knowledge states would be difficult (due to a lack of diagnostic actions, and large hypothesis spaces over potential epistemic states). Our computational framework was based on a growing body of research showing that mental-state inference is structured around an assumption that agents act to maximize utilities—the difference between the costs that agents incur and the rewards they obtain (Jara-Ettinger et al., 2016; Gergely & Csibra, 2003; Lucas et al., 2014; Jern et al., 2017). Our model builds on these ideas, and extends them by explicitly modeling the idea that the costs agents expect to incur often depend on the knowledge they possess. Consequently, others' choices can reveal the costs they appear to be acting under, providing indirect insight into the nature and quantity of their privileged knowledge.

Our results suggest that people can derive graded estimates of how much others know or expect to learn based on limited observable action. Specifically, across two experiments, adults' inferences matched our model predictions in a quantitative manner. Moreover, alternate models which did not consider how knowledge would affect agents' expected costs (and thus their behavior) failed to capture the graded structure of participants' judgments.

Related work has developed computational models that explain how people infer each other's beliefs about the world (Baker et al., 2017) as well as beliefs about their own competence and preferences (Jara-Ettinger et al., 2020). These inferences, however, often depend on access to a limited set of epistemic hypotheses, and to observable behavior that is diagnostic of the agent's epistemic state. While these inferences are undoubtedly critical for social interaction, many everyday social behaviors lack the information needed to make such precise and targeted epistemic inferences. We show that, in such situations, people can nonetheless derive quantitative estimates of how much knowledge someone might possess (or believe they can come to possess).

To make these broad epistemic inferences, our model considered a large space of possible epistemic states and weighted the amount of knowledge expressed in each epistemic hypothesis by its posterior probability (Equation 2.2). This highlights two critical assumptions that our model makes. First, observers must have access to a range of epistemic hypotheses that they can evaluate; second, they must have a way to quantify how much knowledge is expressed in each hypothesis.

While the first assumption may seem plausible in common situations, there are many cases where we do not know what other people's epistemic states could look like (e.g., while we can represent that pilots know how to fly planes, most of us have no idea precisely what a pilot knows). Similarly, the second assumption (that it is possible to quantify the knowledge encoded in each hypothesis) was easy to formalize in the experimental contexts that we considered. But this is not always the case; and, as illustrated by the pilot example, sometimes we may not even know what there is

to know.

These assumptions highlight an important limitation of our findings. Here, our goal was to describe epistemic inference at a computational level (Marr, 1982), not an algorithmic one. It is quite possible participants did not consider the same kinds of hypotheses our model did, and it is unclear how many hypotheses participants sampled before producing a knowledge rating. While our model generated epistemic hypotheses by sampling sets of eggs the agent might know about (Experiment 1) or enumerating the number of island squares whose contents might be known (Experiment 2), it is possible that participants used a completely different algorithm. For instance, adults could learn from experience how knowledge affects their own choices, and draw on previously-learned associations between knowledge and behavior to judge what others know (rather than sampling epistemic hypotheses online). Relatedly, it is an open question whether there are any general measures, such as entropy, that might support quantifying knowledge independent of context. Future work should explore these possibilities.

Our findings raise two additional open questions. First, because each field only contained one prize and each island contained only one treasure chest, it was not necessary for the agent to see inside every egg or search every island square to have full knowledge. Specifically, if the agent knew which egg or island square contained the prize, they could conclude that the rest of the eggs or squares were empty by default. Thus, while the agent's knowledge was often graded (if she didn't know where the prize was, she could still avoid opening eggs or searching squares she knew were empty), it wasn't always (if she knew where the prize was, she could just go to it). Importantly, such situations are commonplace: if you're looking for the silverware drawer in a new kitchen, you can stop searching once you find it—and what's more, you can be fairly confident the other drawers don't contain the silverware you were looking for. However, this is only one situation of many: for instance, if the drawer you found contains only spoons, you might not conclude on this basis that the rest of the drawers do not contain silverware (in fact, you might suspect that if you continue opening drawers, you will find the rest of the set). Thus, knowledge is not always all-or-none, and future work should continue to explore adult epistemic inference in a wider variety of contexts.

Finally, we proposed that our task elicited, and model captured, amorphous knowledge inferences. However, we did not test whether participants can also make specific inferences in this task (judging how likely the agent is to know the contents of each egg). While many different knowledge states could have given rise to each observed choice, it is still true that in Experiment 1, knowledge over certain eggs affects the expected utility of an action plan more than others. For instance, we can be fairly confident that the agent in Figure 2.1, Panel A knows that the furthest eggs in field A are empty—otherwise, the expected cost of locating the prize in this field could be quite high. However, it is not as clear whether the agent knows the contents of the more central eggs, and it is difficult to distinguish whether the agent is more likely

to know about some than others (because knowledge over each affects the agent's expected costs similarly). Our model is already able to infer the agent's average knowledge over each egg in a scenario; future work will test whether our model can also capture any specific judgments participants can make—testing whether participants make strong inferences where our model does, and are uncertain when our model is.

2.5.1 Conclusion

In the current work, we explain a common everyday epistemic inference: the ability to infer how much others know or believe they can learn, even when specific epistemic inferences are under-determined. But as we navigate the world, these are not the only epistemic inferences we make. From observing the outcome of others' goal-directed actions we infer what they thought they knew (and what they actually did); from observing their choices to seek information (and at what cost) we infer how much they believed they could learn; and from observing their interactions with other agents, we infer how much they think others know. The work presented here serves as an initial step towards modeling the full scope of epistemic inferences people make in their everyday lives.

2.6 Acknowledgments

We thank the members of the Yale Computational Social Cognition Lab for helpful conversations and advice. This work was supported by NSF award BSC-2045778, and also by Yale's Franke Program in Science and the Humanities, via a Franke Interdisciplinary Graduate Award to RA.

Chapter 3

Do preschoolers rely on a Theory of Mind to make epistemic inferences?

This chapter is based upon Aboody, Huey & Jara-Ettinger (under revision). Preschoolers decide who is knowledgeable, who to inform, and who to trust via a causal understanding of how knowledge relates to action

Abstract

Preschoolers are discerning learners, preferring to trust people who are accurate, reliable, and appropriately-informed. Do these preferences reflect mental-state reasoning, where children infer what others know from their behavior, or do they reflect a reliance on simple cues? In Experiment 1 we show that four- and five-year-olds can infer knowledge from others' behavior when superficial cues and actions are matched across agents. Experiments 2a and 2b further show that children track how agents acquired their knowledge, and use this to determine what different agents will (and will not) know. Finally, Experiment 3 shows that children spontaneously make these knowledge inferences when deciding whom to trust. Our findings suggest that by age four, children have expectations about how knowledge relates to action, use these expectations to infer what others know from what they do, and rely on these inferences when deciding whom to trust.

3.1 Introduction

Most of what we know, we learn from others. We learn what foods are safe to eat without poisoning ourselves, compute derivatives without rediscovering calculus, and find out what's happening in Washington without witnessing it firsthand. Yet, while social learning removes the cost and risk of exploration, it introduces a new challenge: How do we know whether others' testimony is true? Agents can be misinformed, unreliable, or malicious, and it is critical to distinguish those who are knowledgeable and trustworthy from those who are not.

The ability to identify useful informants is rooted early in development. By preschool, children selectively learn from those who are accurate (preferring agents who label objects correctly; Koenig, Clément, & Harris, 2004; Koenig & Harris, 2005), reliable (preferring agents who behave consistently over time; Chow, Poulin-Dubois, & Lewis, 2008; Zmyj, Buttelmann, Carpenter, & Daum, 2010), and adequately-informed (preferring agents who had access to relevant information; Koenig, 2012; Robinson, Butterfill, & Nurmsoo, 2011). These and related findings (see Sobel & Kushnir, 2013 for review) raise the possibility that children use a causal mental model of how knowledge relates to action—a Theory of Mind—to navigate the epistemic world.

However, children's preference for accurate, reliable, and well-informed agents may not be grounded in mental state reasoning. Preschoolers struggle to reason about beliefs explicitly (Wellman et al., 2001; H. Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018), and approximating mental-state inferences through simpler rules might be easier and generally effective. Indeed, children often rely on superficial cues that do not necessarily reveal agents' knowledge, preferentially learning from those who are familiar, (Corriveau & Harris, 2009a), dominant (Bernard et al., 2016), and those who share their accent (Kinzler, Corriveau, & Harris, 2011).

Here we focus on three open questions. First, to what extent do children select informants based on superficial behavioral cues? Do children circumvent mental-state inferences by associating different epistemic states with different observable cues (e.g., possibly equating knowledge with accuracy, and ignorance with error; Ruffman, 1996; although see Friedman & Petrashek, 2009)? Or can children infer knowledge through their Theory of Mind (e.g., considering which knowledge states best explain an observed set of actions)? Second, what are children's attributions of knowledge like? Are they broad attributions (e.g., general beliefs like "Max is knowledgeable," as suggested by halo-like effects in children; Brosseau-Liard & Birch, 2010) or representations that track the content of what others know (e.g., concrete beliefs like "Max is knowledgeable about what's inside this box")? Finally, do children derive such inferences spontaneously when necessary, or only when explicitly prompted?

We present three studies that shed light on these questions, using simple social events where two agents take identical actions and produce identical outcomes, differing only in the order in which they produce each action. This enabled us to design events where agents are indistinguishable in terms of superficial cues, but whose behavior reveals different epistemic states when analyzed through a Theory of Mind.

We focus on four- and five-year-olds because younger children struggle to explicitly reason about other people's beliefs and knowledge (Robinson & Whitcombe, 2003; Wellman et al., 2001).

3.2 Approach to analyses

Following current recommendations for statistical best practices, we take an estimation approach to data analysis (as opposed to relying on null-hypothesis significance testing; Cohen, 1994; Cumming, 2014). We estimate effect sizes by bootstrapping our data and obtaining 95% confidence intervals; we take confidence intervals that do not cross chance as evidence of a reliable effect. To estimate the difference in effects between two groups, we obtain 95% difference confidence intervals over the difference between groups. We take confidence intervals that do not cross 0 (the point of no difference) as evidence of a reliable effect.

3.3 General methods

The procedure, analysis plan, sample size, and exclusion criteria for all experiments were pre-registered, unless otherwise indicated. All pre-registrations, scripts, data, and analysis files are available in the OSF project page: `https://osf.io/mhcvt/?view_only=5445c5172f1f41768f73f7a2812425f9`. The pre-registered sample size for all experiments was determined through Monte Carlo power analyses. A post-hoc analysis confirmed that our sample was appropriately powered, with power > 0.8 for every test question in every experiment (see Supplemental Materials for details: `https://osf.io/x26ha/`).

In Experiments 1 and 2a, participants were asked two test questions, order counterbalanced. As we collected data for Experiments 1 and 2a (run approximately concurrently), we realized that if we obtained an age difference, $n = 16$ participants per age group might not yield sufficient power to detect potential order effects. Thus, even before we finished collecting this full dataset, we chose to double our sample, replicating each experiment. For both experiments, performance did not differ between samples (see Supplemental Materials). We report both analyses on each sample separately, and also aggregate this data to compute effect sizes more accurately—but do not compute p-values.

3.4 Experiment 1

Experiment 1 tests whether preschoolers can infer which of two agents is knowledgeable when agents are matched for low-level superficial cues. Participants were introduced to two puppets, one that accurately predicted what was under two cups before revealing their contents; and one that accurately stated what was under the

cups after revealing their contents. If children rely on simple behavioral cues like accuracy or checking, they should believe that both agents are equally knowledgeable. But if children consider the epistemic states that best explain each agent's actions, they should judge that the predictor is more likely to have been knowledgeable.

3.4.1 Method

Participants

64 four- and five-year-olds (mean age: 5.02 years, range: 4.07 – 5.99 years; n = 32 participants per age group) were recruited at preschools (n = 4), children's museums (n = 57), or in-lab (n = 3). 32 participated in the pre-registered original experiment. Because we asked two test questions, we became uncertain whether our original sample would be sufficient to account for potential order effects—thus, we chose to double our initial sample, collecting an additional 32 participants in a direct replication. The direct replication was not pre-registered but was identical in all aspects to the original pre-registered experiment. 23 additional participants were recruited but not included in the study (21 at museums, and 2 at preschools; see Results).

Stimuli

Stimuli consisted of two male puppets, three paper cups (red, blue and yellow), and three small animal figurines (a fox, a hippo, and a deer).

Procedure

The participant and the experimenter sat on opposite sides of a table. The experimenter first showed participants three cups (sitting inverted on the table), and lifted each cup to reveal an animal figurine hidden underneath. The experimenter then introduced two puppets, Max and Sam, explaining that, "Right before you came in here, one of our friends snuck out from under the table and peeked underneath all the cups! And one friend stayed under the table, and he never saw anything." The experimenter then explained that she'd "ask our friends some questions to find out who peeked underneath all the cups."

Participants were allowed to choose which puppet they wanted to hear from first, and which cup they wanted to ask the puppet about. This was to avoid any pragmatic concerns that could arise if participants interpreted the experimenter's choice of puppet or cup as meaningful (e.g., perhaps the experimenter suspects the puppet she questions first?). Puppets' roles (predictor/observer) were assigned after participants chose which puppet to ask first, so the role assigned to the first puppet was counterbalanced across participants.

When the predicting agent was asked what was under a cup, he correctly stated the animal name, lifted the cup to reveal its contents, and then looked at the animal (e.g., saying 'There's a hippo under this cup,' revealing a toy hippo, and looking

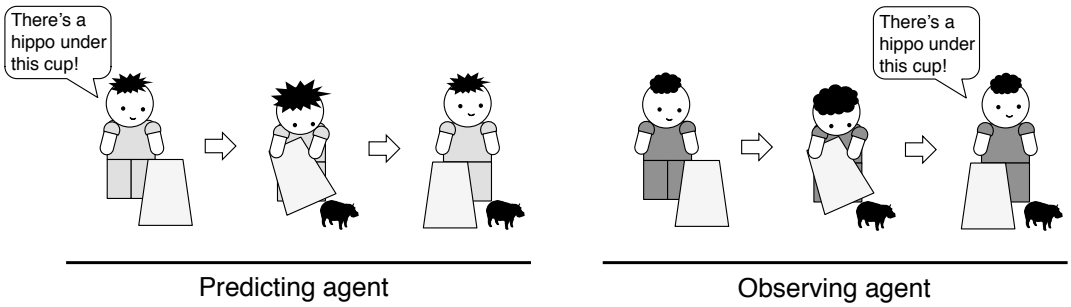

Figure 3.1: Agents' behavior in Experiments 1-3. The predicting agent first stated the contents of the cup, and then revealed them. The observing agent first revealed the contents of the same cup, and then described them. In Experiment 1, participants were asked which puppet had peeked under the cups before the task began, and which puppet knew what was under the final cup, which neither puppet had interacted with. In Experiments 2a and 2b, the animal under the final cup was replaced without the puppets' knowledge. In Experiment 2a participants were asked the same questions as in Experiment 1. In Experiment 2b participants were again asked which puppet had peeked, and they were given an opportunity to tell one of the puppets the name of the animal currently under the cup. Finally, in Experiment 3, participants watched the puppets disagree about the contents of the final cup, and were asked to endorse one agent's testimony.

down at it; Figure 3.1). By contrast, when the observing agent was asked what was under a cup, he first lifted the cup to reveal its contents, looked down, and then correctly stated the animal name (e.g. revealing a toy hippo, looking down at it, and then saying 'There's a hippo under this cup'). Thus, both puppets performed identical actions, but in the opposite order: the predictor first said what was under the cup and then looked; the observer first looked under the cup, and then said what was there (see Figure 3.1). After each puppet stated the contents of the first cup, participants chose a second cup to ask the puppets about and each puppet repeated the same actions (predicting or observing the cup's contents), in the same order as in the first trial. Each puppet was always alone when asked what was under each cup.

Participants then answered two test questions, order counterbalanced. In the prior knowledge test question, participants were asked, "which one of our friends peeked?" And in the generalization test question, participants were asked, "which one of our friends knows what's under this cup?" (referring to the last remaining cup, which neither of the puppets had interacted with); see Supplemental Materials for scripts and additional details.

3.4.2 Results

For the 94% of participants whose sessions were video or audio taped (n = 82/87), two coders who were not involved in data collection determined exclusions according to

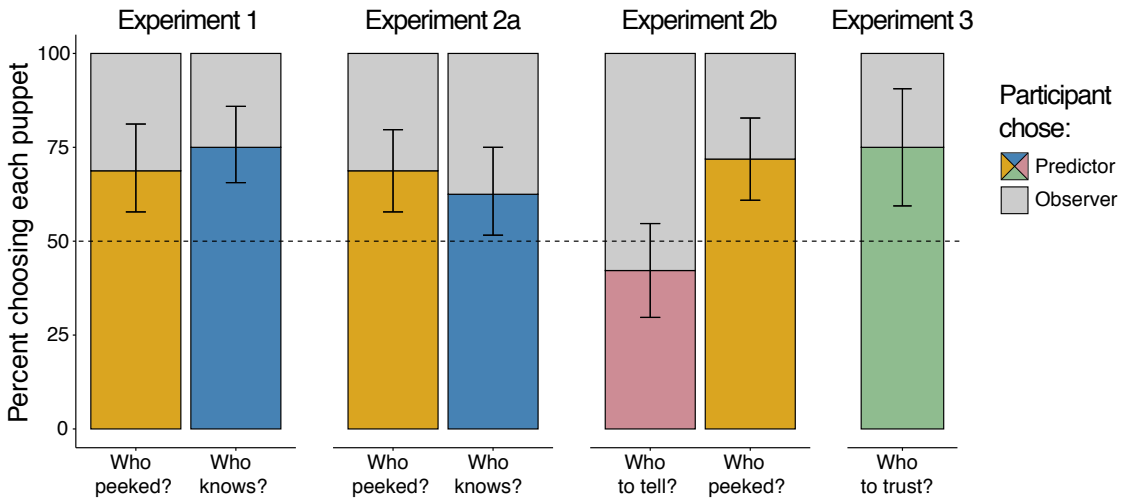

Figure 3.2: Results from all three experiments. The dotted line indicates chance performance, and the error bars are bootstrapped 95% confidence intervals. The brightly-colored portion of each bar shows the proportion of participants who selected the predictor, with each color corresponding to a different test question. The gray portion of each bar shows the proportion of participants who selected the observer. In Experiment 1, participants judged that the predicting agent had peeked underneath the cups, and knew what was under the remaining cup. In Experiment 2a, participants judged that the predicting agent had peeked under the cups, but also judged that he knew what was under the remaining cup (whose contents had been switched out). In Experiment 2b, participants again judged that the predicting agent had peeked, and had no preference when given a chance to tell an agent the name of the new animal that had been placed under the last cup. In Experiment 3, participants preferred to endorse the predicting agent's testimony when the two agents disagreed about the contents of the last cup, suggesting that they inferred who was knowledgeable without any explicit prompts.

pre-registered criteria. The first coder was blind to participants' final answers, checking for any experimenter errors, family interference, and ensuring that the participant was attentive. The second coder, blind to condition, checked whether participants answered the test questions, and whether the experimenter or family members behaved in any way that could affect their choice. For participants who were not video or audio taped (6% of participants; n = 5/87), the experimenter took notes on any deviations from the procedure, and the first author determined exclusions by comparing these notes to the pre-registered inclusion criteria. Twenty-three participants were recruited but not included in the study, because the

participant interfered with the study (by revealing the contents of the cups to the puppets; n = 11), due to experimenter error (n = 6), because the participant did not complete the study (n = 4), or because they did not speak English fluently (n = 2).

Of the final 64 participants included in the study, 68.8% judged that the predicting agent had peeked under the cups (n = 44 of 64; 95% CI: 57.8 – 81.2) and 75% judged that the same agent also knew what was inside the last cup (n = 48 of 64; 95% CI:

65.6 – 85.9). See Figure 3.2. The proportion of participants selecting the predicting agent in response to each test question did not differ (95% difference CI: -9.4 – 21.9). A Bayesian mixed-effects logistic regression revealed no effects of age, question type, or their interaction (all $|\beta|$'s between [0.23, 0.72]; with all 95% equal-tailed credible intervals crossing 0 (indicating the absence of a reliable effect; not pre-registered).

These results are qualitatively unchanged if performance is analyzed separately in each sample, with participants in both samples reliably selecting the predicting agent in response to both test questions. However, note that a post-hoc power analysis suggests that separating the data by sample leaves us under-powered to detect effects in most of our comparisons; see Supplemental Materials. In the original sample, a Bayesian mixed-effects logistic regression (not pre-registered) revealed no effects of age, question type, or their interaction. In the replication sample the same regression did reveal an interaction between age and test question—but given that we obtained no such effect when aggregating our data, or in a meta-analysis over all experiments (see Supplemental Materials), and given our low power when separating the samples, there is no indication such an interaction would be reliable in a more highly-powered sample. See Supplemental Materials for full results, supplementary analyses, and regression tables.

3.5 Experiment 2a

Experiment 1 shows that children can infer which of two agents is knowledgeable in the absence of superficial cues. Because both agents behaved identically (varying only in the order of their actions), our results begin to go beyond past work, which has shown that children realize agents who simply repeat others' accurate testimony (Einav & Robinson, 2011) or ask accurate questions (Luchkina, Sobel, & Morgan, 2018; Luchkina, Morgan, Williams, & Sobel, 2020) are not necessarily knowledgeable.

In Experiment 2 we explore the representational nature of these epistemic attributions. Do children's inferences result in coarse, broad knowledge attributions? Or do children use an agent's behavior not only to infer if they are knowledgeable, but to determine the exact contents and limits of what they know?

3.5.1 Method

Participants

64 four- and five-year-olds (mean age: 4.97 years, range: 4.0 – 5.99 years; n = 32 participants per age group; 32 in the pre-registered original experiment, and 32 in a direct replication) were recruited at preschools (n = 8), children's museums (n = 52), or in-lab (n = 4). Because we asked two test questions, we became uncertain whether our original sample would be sufficient to account for potential order effects—thus, we chose to double our initial sample, collecting an additional 32 participants in a direct replication. The direct replication was not pre-registered; it is identical in all aspects

to the original pre-registered experiment. 21 additional participants were recruited but not included in the study (14 from children's museums, 2 from preschools, and 5 from festivals in the New Haven area; see Results).

Stimuli

Materials were identical to those of Experiment 1, with the addition of a small box (6 x 7 x 7 in.) containing six animal figurines: a cat, a duck, a penguin, a parakeet, a rabbit, and an ostrich.

Procedure

The procedure was identical to that of Experiment 1 with one exception. After the two puppets predicted or observed the contents of the first two cups, the puppets left and the experimenter said, "But you know what? We haven't asked our friends about this cup yet," and pointed to the remaining cup. The experimenter brought out a box and continued, "And I thought we could play a trick. I have this box of animals—can you choose one?" After the participant chose a new animal figurine, the experimenter said, "Ok! So let's put [original animal] back in this box, and let's put [new animal] underneath. So now, [new animal] is here instead!" The puppets were then brought back, and participants answered the same test questions (prior knowledge and generalization; order counterbalanced); see Supplemental Materials for scripts and additional details.

3.5.2 Results

Results were coded as in Experiment 1, according to identical pre-registered criteria. 89.4% of participants were video or audio taped (n = 76/85). 21 participants were recruited but not included in the study because the participant interfered with the study (by revealing the contents of the cups to the puppets; n = 12), due to experimenter error (n = 3), because the participant was distracted (n = 2), non-neurotypical (n = 2), did not answer a test question (n = 1), or due to sibling interference (n = 1).

If children attribute epistemic states based on a causal understanding of how knowledge relates to action, they should continue to judge that the predictor peeked. However, they should no longer judge that he would also know what was underneath the last cup—because neither agent had seen the switch—performing at chance in the generalization question. Consistent with our first prediction, 68.8% of participants judged that the predicting agent had peeked under the cups (n = 44 of 64; 95% CI: 57.8 – 79.7), replicating our finding in Experiment 1. Contrary to our predictions, however, 62.5% of participants judged that the predicting agent also knew what was under the last cup, a proportion reliably higher than chance (n = 40 of 64; 95% CI: 51.6 – 75). The proportion of participants selecting the predicting agent in response

to each test question did not differ (95% difference CI: -21.9 – 10.9). Moreover, the proportion of children judging that the predicting agent knew what was under the last cup was not reliably lower than the proportion we found in Experiment 1 (75% of participants in Experiment 1 compared to 62.5%; 95% CI: -28.1 – 3.1). A Bayesian mixed-effects logistic regression revealed no effects of age, question type, or their interaction (all $|\beta|$'s between [0.30, 0.34]; with all 95% equal-tailed credible intervals crossing 0; not pre-registered).

These results are qualitatively unchanged if performance is analyzed separately in each sample, with the exception that in the replication sample, participants had no reliable preference for either agent when asked who knew what was under the last cup (n = 19 of 32; 59.4%; 95% CI: 43.8 – 78.1). Although this is the only case across both Experiment 1 and Experiment 2a that judgments in any sample did not reliably differ from chance, note that as in our prior experiment, a post-hoc power analysis suggests that separating the data by sample leaves us under-powered to detect effects in most of our comparisons; see Supplemental Materials. A Bayesian mixed-effects logistic regression (not pre-registered) revealed no effects of age, question type, or their interaction in either sample. See Supplemental Materials for full results, supplementary analyses, and regression tables.

Why did participants judge that the predictor knew what was under the final cup when the animal figurine had just been replaced? One possibility is that children inferred that the agent's accurate prediction implied he had peeked and was therefore knowledgeable, but then failed to recognize the boundaries imposed by this inference. Alternatively, however, participants may have believed that neither puppet knew what was under the last cup, but defaulted to the predictor for lack of a better option (after all, the predicting agent knew what was under the final cup at one point in time, while the observing agent never did). To disentangle these possibilities, we ran an exploratory analysis testing whether children's response times varied as a function of test question. A coder blind to our hypotheses recorded participants' response times to the test questions in Experiments 1 and 2a (for all participants where audio or video was available; 95.3% and 89.1% of participants in Experiments 1 and 2a, respectively), measuring the time between the end of the test question and the onset of the participant's answer.

Children's response times in Experiment 1 were comparable across questions, and did not significantly differ (prior knowledge test question, M = 2.56s, SD = 3.64; generalization test question, M = 3.20s, SD = 4.25; t(60) = -0.96, p = 0.34, d = 0.12 by paired t-test). By contrast, children's response times in Experiment 2a were significantly different, with participants requiring an average of 2.18 additional seconds to answer the generalization question relative to the time it took them to answer the prior knowledge test question (prior knowledge question, M = 1.73s, SD = 2.17; generalization question, M = 3.91s, SD = 4.83; t(56) = -3.13, p = 0.0028, d = 0.42 by paired t-test). See Figure 3.3.

These exploratory analyses provide some initial evidence that participants were

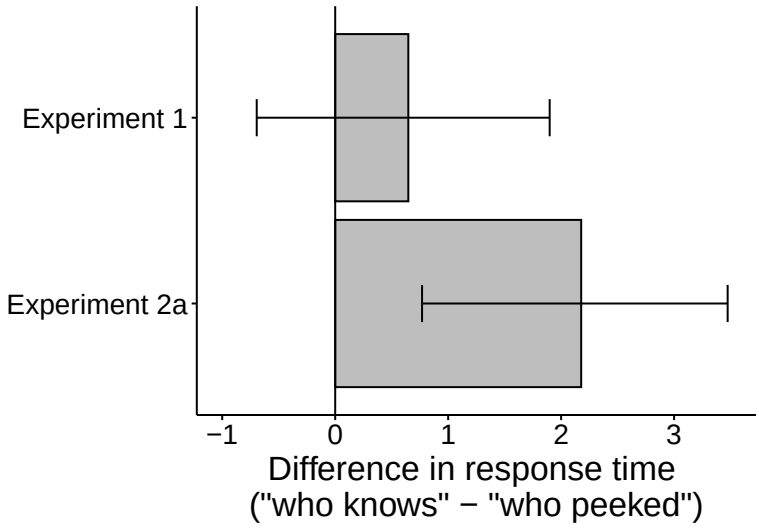

Figure 3.3: The difference in response time ("who knows" – "who peeked") in seconds for Experiments 1 and 2a. Error bars are paired 95% bootstrapped confidence intervals over the difference in response time. In Experiment 1, there is no reliable difference in reaction time across the two test questions (the confidence interval crosses 0, the point of no difference). In Experiment 2, there is a reliable difference, with the same participants taking reliably longer to answer the "who knows" generalization test question than they took to answer the "who peeked" prior knowledge test question.

uncertain about how to answer the generalization question in Experiment 2a, opening the possibility that children indeed recognized that neither agent knew what was under the last cup, but defaulted to the predicting agent due to the forced-choice nature of our task. Conversely, given that children seemed to answer both test questions without problem in Experiment 1, participants in Experiment 1 may have indeed answered the generalization question by considering how the predictor gained his knowledge (rather than hesitantly defaulting to this agent for lack of a better answer).

3.6 Experiment 2b

Children's choices in Experiment 2a suggest that they did not track the limits agents' knowledge, but their response times point to the possibility that they did. To distinguish between these possibilities, Experiment 2b replicated Experiment 2a; but rather than asking children who knew what was under the final cup, participants were given the chance to inform one of the puppets. If participants believe that the predictor knows what's under this cup (as suggested by their answers in Experiment 2a), they should prefer to inform the observer. However, if participants believe that both agents are now ignorant (as suggested by their response time in Experiment 2a), they should respond at chance.

3.6.1 Method

Data collection for this study was interrupted due to the COVID-19 pandemic and this task was therefore completed online. To ensure that the switch to online testing did not affect our results, we first conducted a validation pilot, which revealed no differences between in-person and online samples (see Supplemental Materials). The revised plan in response to the pandemic was pre-registered, and both the original and revised pre-registration are available in our OSF repository.

Participants

To best compare Experiment 2b to the prior experiments, we pre-registered a sample of 64 participants (n = 32 per age group). Thus, 64 four- and five-year-olds (mean age: 5.06 years, range: 4.0 – 5.91 years; n = 32 participants per age group) participated. The first 15 were recruited in-person, at children's museums (n = 14) and preschools (n = 1); the final 49 were recruited online. 8 additional participants were recruited but not included in the study (2 from preschools, 1 from a children's museum, and 5 recruited online; see Results).

Stimuli

For the in-person procedure, materials were identical to those of Experiment 2a. For online participants, the experimenter narrated while playing a pre-recorded puppet show (embedded in a PowerPoint presentation). The script and pre-recorded show were kept as similar as possible to the in-person version. The PowerPoint presentation is available in the project OSF page.

Procedure

The original in-person experiment began and proceeded in the same way as Experiment 2a up to the test question phase of the task. At this point, instead of asking participants who would know what was underneath the final cup, the experimenter offered participants the chance to tell one of the puppets what was there, saying, "So, we haven't asked our friends about this cup yet [pointing to cup]. But guess what! You can tell one of our friends the name of the animal underneath this cup. Which one of our friends do you want to tell?" Participants in this task were always asked who they wanted to inform before being asked which puppet had peeked. This pre-registered decision helped us avoid any potential influence of the peeking question on children's decision of who to inform (this was a conservative decision: we did not observe any order effects in our previous experiments, with the exception that in Experiment 1, participants were more likely to correctly answer the "who peeked" prior knowledge question when it was asked first).

The online task was designed to be as close as possible to the in-person design, with three pre-registered procedural adjustments. First, participants were no longer

allowed to choose the puppet they wanted to question first, and the cups they wanted to ask about, as this was impractical for online testing given that the puppet show was prerecorded. Each puppet's role as predictor or observer was still counter-balanced across participants. Second, because participants could not reach in and choose an animal to place under the final cup, the actor in the video chose one randomly. To achieve this, the actor in the video placed the box of animals on the table, and then upended it, spreading out all of the animals on the table. The experimenter narrated, "I have this box of animals. Look! Do you see the animals? Look!" After participants had a chance to see the animals, the experimenter continued, "Let's put them back in the box. And we'll close our eyes and choose one." The actor replaced the animals in the box, shook it around for several seconds, reached in, sampled an animal, and then held it out to the camera to show that she had pulled out a cat. The experimenter narrated, "Look! We got a kitty cat!" As the actor replaced the contents of the blue cup, the experimenter said, "Ok! So let's put the hippo back in this box, and let's put the kitty cat underneath. So now the kitty cat is here instead!"

Finally, each puppet was associated with a color (blue and green, indicated by their shirt) so that children could select a puppet by choosing a color (in line with recent work showing this to be an effective method for online testing of four- and five-year-olds; Aboody et al., 2021a; Kominsky et al., 2021). In addition, the experimenter emphasized their shirt color when the puppets were introduced, saying, "This is Sam, wearing a green shirt. And this is Max, wearing a blue shirt." Similarly, when asking the test questions, the experimenter emphasized the options by referring to each agent's shirt color. For the tell test question, the experimenter said, "So, we haven't asked our friends about this cup yet. But guess what! You can tell one of our friends the name of the animal underneath this cup. You can tell our friend in the blue shirt, or you can tell our friend in the green shirt. Which one of our friends do you want to tell? The one in the blue shirt, or the one in the green shirt?" And for the prior knowledge test question, the experimenter said, "And can you tell me: which one of our friends peeked? The one in the blue shirt, or the one in the green shirt?"

3.6.2 Results

Results were coded in the same way as Experiments 1 and 2a, according to identical pre-registered exclusion criteria (but with the addition that coders would also check for any internet connectivity issues). 87.5% of participants were video or audio taped (n = 63/72). 8 participants were recruited but not included in the study due to experimenter error (n = 4), because the participant interfered with the study (by revealing the contents of the cups to the puppets; n = 1 in-person participant), because they did not complete the experiment (n = 1), because they did not answer a test question within 30s (n = 1), or because they had already participated in the past (n = 1).

When asked whom they wanted to inform, 42.2% of participants selected the predicting puppet, and 57.8% selected the observing puppet. These proportions are

not reliably different from chance (n = 27 of 64 chose to inform the predicting puppet; 95% CI: 29.7 – 54.7). This chance performance, however, was not due to a failure to track or understand the experiment: 73.4% of participants judged that the predicting puppet had peeked (n = 47 of 64; 95% CI: 62.5 – 84.4), replicating our findings from Experiments 1 and 2a. These results suggest that children's inference that the predictor had peeked and was therefore knowledgeable also allowed them to track the limits of what the predictor would know.

A Bayesian mixed-effects logistic regression (not pre-registered) found no main effects for age (β = -0.71, 95% CI: -1.74 – 0.28), but revealed a main effect for question type (with participants more likely to select the predictor as the one who peeked vs. the one they would like to inform; $\beta = 1.5$, 95% CI: 0.72 – 2.30). Finally, a Bayesian mixed-effects logistic regression (not pre-registered) found no main effect of testing location (in-person vs. online; $\beta = 1.06$, 95% CI: -0.33 – 2.57), but revealed an interaction between testing location and test question (with online participants more likely to select the predictor when asked who they would like to inform, β = 2.40, 95% CI: 0.25 – 4.90, and less likely to select the predictor when asked who peeked, β = -2.39, 95% CI: -4.83 – -0.27; see Supplemental Materials for regression tables). Note however that only 15 of our 64 final participants were run in-person; thus, it is unclear whether these differences would be reliable given a larger in-person sample. Furthermore, although in-person participants were better able to answer the "who peeked" test question, the majority of our participants were run online, and overall participants still reliably identified the predictor as the one who peeked. Finally, even if participants were truly better able to track the limits of the puppets' knowledge in our online task (although again, it is unclear whether this would be the case given a larger in-person sample), this would still serve as evidence that young children are capable of doing so.

3.7 Experiment 3

Experiments 1-2b suggest that children can infer knowledge from agents' actions in the absence of superficial cues, and that their epistemic attributions track what agents do or do not know. However, participants in both experiments were asked to make epistemic judgments explicitly. In Experiment 3, we test whether these capacities underlie preschoolers' decisions about whom to trust when two agents provide conflicting testimony.

3.7.1 Method

Participants

Because we asked only one test question in Experiment 3, it was not necessary to increase our sample to account for potential test question order effects. Thus, we pre-registered and recruited a single sample of 32 four- and five-year-olds (mean age:

5.02 years, range: 4.14 – 5.99 years; n = 16 participants per age group). These participants were all recruited at a children's museum. Seven additional participants were recruited but not included in the study (these participants were also recruited at children's museum; see Results).

Stimuli

Materials were identical to those of Experiment 1.

Procedure

The task began in the same way as Experiment 1 with the difference that participants were not shown the contents of the cups at the beginning of the task. After the two puppets had interacted with the first two cups, the experimenter pointed to the third cup, and said, "Well, we didn't ask our friends about this cup yet. So let's ask both of our friends about what's under this cup." One puppet said, "There's a bear under this cup," and one said "There's a squirrel under this cup" (randomizing which puppet spoke first, and the animal they claimed was under the cup). Finally, participants were asked "Can you tell me: what animal is under this cup?" After the test question, participants were asked the same memory check questions from Experiment 1 (order randomized).

3.7.2 Results

Results were coded in the same way as Experiments 1-2b according to identical pre-registered criteria. 97% of participants were video or audio taped (n = 38/39). Seven participants were recruited but not included in the study because the participant did not answer the test question (despite prompting; n = 5), due to experimenter error (n = 1), or because the participant lifted a cup and revealed its contents during the puppet show (n = 1). 75% of participants endorsed the predicting agent's testimony (24 of 32; 95% CI: 59.4 – 90.6), suggesting that children spontaneously used the agent's behavior to infer knowledge, even when they were not explicitly prompted to do so. See Figure 3.2. A Bayesian mixed-effects logistic regression revealed no effect of age (β = 0.54, 95% CI: -0.87 – 1.98; not pre-registered; see Supplemental Materials).

3.8 General Discussion

The capacity to determine and track what people know is critical for social life, from everyday communication (Bohn & Köymen, 2018) to social learning (Harris, 2012). Here we explored the representations and inferences that preschoolers use to navigate the social world, focusing on three questions. First, do preschoolers attribute

knowledge through a sensitivity to simple cues (such as assuming that accuracy inevitably implies knowledge) or via mental-state reasoning, where children consider what epistemic states best explain someone's observed behavior? Second, when children attribute knowledge, do these representations include expectations of what an agent may or may not know? Or are they coarser attributions? Finally, do children make such inferences spontaneously when deciding whom to trust? To test these questions, we presented children with simple events where two agents took identical actions to produce identical outcomes. By varying the order of agents' actions, we were able to create situations where children could distinguish the agents' epistemic states only if they reasoned about which mental states best explained each agent's behavior.

Experiment 1 showed that four- and five-year-olds distinguish between accurate predictions and accurate observations, and use this distinction to infer agents' causal history (who peeked?) and epistemic states (who knows what's under a new cup?). Experiments 2a and 2b showed that children's beliefs about how agents gained their knowledge allow them to determine the limits of what an agent knows. Finally, Experiment 3 showed that children make these inferences spontaneously when deciding whom to trust (inferring that the predictor was knowledgeable from his actions), and implicitly use this epistemic inference when deciding whom to believe (without being asked to reason explicitly about knowledge).

Could children in our task have used a simple heuristic that associates accurate predictions and knowledge, without relying on intuitions of how mental states relate to action? While such a heuristic could be a sensible rule to rely on in many contexts (those who are accurate are often indeed knowledgeable), Experiment 2b suggests that children can move beyond such an association when needed. Specifically, in Experiment 2b children were equally likely to inform each puppet, showing that they did not superficially associate accuracy with knowledge (if they had, children ought to have informed the puppet who made observations). Furthermore, identifying a statement as a prediction already requires children to reason about the connection between knowledge and action: children must track what information was available to each agent before they made a statement, and consider whether this information could explain their accuracy. Therefore, even if children succeeded in the task by associating accurate predictions to knowledge, children's ability to conceptualize agents as making predictions or observations would already be evidence of a reliance on a causal mental model of how knowledge relates to action.

Our results are consistent with related work showing that, by age four, children recognize that agents who repeat facts that they heard from others, or agents that simply ask accurate questions, are not necessarily knowledgeable (Einav & Robinson, 2011; Luchkina et al., 2018, 2020). This work shows that children disregard accuracy when it doesn't emerge from an independently-produced statement (i.e., stated declaratively rather than as a question, and produced without help from others). Our work goes beyond previous research by testing whether such behavior reflects a

causal understanding of how knowledge relates to accuracy, or a simpler belief that independently-produced accurate statements imply knowledge. Our work contributes to this literature by providing evidence that children's inferences are sensitive to subtle differences in behavior that can reveal agents' knowledge when analyzed through a causal mental model of others' epistemic states.

At the same time, our results do not imply that children never rely on heuristics when reasoning about knowledge. Related work shows that children select informants based on a variety of cues, such as agents' accent, familiarity, and dominance (see Introduction); young children will even favor familiar agents over accurate ones (Corriveau & Harris, 2009a). This work opens three possibilities. A first possibility is that young children are simultaneously motivated to learn from knowledgeable agents and to affiliate with in-group members (Dunham, Baron, & Banaji, 2008). If so, children's apparent reliance on heuristics might reflect intergroup preferences rather than epistemic reasoning. A second possibility is that children do believe that seemingly irrelevant features, such as familiarity or accent, reveal whom to trust (perhaps because familiar people have had a long track record of generally being accurate, or because people with similar accents might have information that is relevant to their group; Begus, Gliga, & Southgate, 2016). Finally, a third possibility is that children begin to navigate the epistemic world by relying on simple heuristics. Such heuristics may be replaced or complemented by richer mental-state reasoning as children's Theory of Mind develops. From this standpoint, four- and five-year-olds may be at an intermediate stage of development where they need not necessarily rely on simple heuristics (vs. mental-state reasoning), but still rely on some superficial cues when these are available. It is also possible that, through mental state reasoning, children develop more effective shortcuts or rules to rely on (for instance, codifying intuitions grounded in Theory of Mind, or identifying the most efficient heuristics for different tasks; e.g., Horn, Ruggeri, & Pachur, 2016)—but still relying on their causal model of other minds to attribute knowledge in situations where simple rules cannot do. These are questions we hope to explore in future work.

Our work also opens a new question. What are the underlying computations that led children to infer that the predictor was knowledgeable? While this is an open question, our results can be readily explained by expanding current theories of mental-state inference. Research suggests that children and adults infer mental states through an assumption that agents maximize their subjective utilities—the difference between the cost they incur and the reward they obtain (Aboody, Denison, & Jara-Ettinger, 2021; Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; Jern et al., 2017). Our findings are consistent with this framework: When the goal is to provide accurate information, a knowledgeable agent can maximize their utilities by providing the correct answer (and thus incurring no unnecessary costs). An ignorant agent, by contrast, ought to incur a cost to obtain the information needed to fulfill their goal (provided that the cost of getting the information does not outweigh the reward associated with getting things right). Note, however, that in our task, the

predictor lifted the cup after they stated what was inside it. Under this framework, this additional cost can be interpreted as evidence that the predictor's goal was not only to provide accurate information, but also to prove that the information was accurate. While more research is needed to test this prediction in children, in an additional study we have shown that adults indeed interpret the observer's cup-lifting as checking (i.e., providing information for the self) and the predictor's cup-lifting as showing (i.e., providing information to others; see Supplemental Materials).

3.8.1 Conclusion

At an age where we gain knowledge primarily by learning from others, our results show that children do not select informants through coarse rules or heuristics. Instead, children navigate the epistemic world by observing others' behavior, and reasoning about the events and mental states likely to have caused that behavior. Our work highlights children's early inferential capacities, and advances our understanding of how these inferences work within children's developing Theory of Mind.

3.8.2 Acknowledgments

We thank the Boston Children's Museum, the Peabody Natural History Museum, and the families who participated in this research. We thank Colin Jacobs for assistance with data collection. We thank Madison Flowers, Caiqin Zhou, Colin Jacobs, Gwyneth Heuser, Eleanor Iskander, Lauren Barragan, and Jenna Landy for help with coding. This work was supported by the Simons Center for the Social Brain. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

Chapter 4

Do preschoolers expect others to maximize their epistemic utilities?

This chapter is based upon Aboody, Zhou & Jara-Ettinger (2021). In Pursuit of Knowledge: Preschoolers Expect Agents to Weigh Information Gain and Information's Cost When Deciding Whether to Explore. *Child Development*.

Abstract

When deciding whether to explore, agents must consider both their need for information and the cost of obtaining it. Do children recognize that exploration reflects a trade-off between action costs and expected information gain, and can they infer epistemic states accordingly? In two experiments, four- and five-year-olds (N=144) judge that an agent who refuses to obtain low-cost information must have already known it, and an agent who incurs a greater cost to gain information must have a greater epistemic desire. Two control studies suggest that these findings cannot be explained by low-level associations between competence and knowledge. Our results suggest that preschoolers' Theory of Mind includes expectations about how costs interact with epistemic desires and states to produce exploratory action.

4.1 Introduction

Humans are intrinsically motivated to learn about the world (Kidd & Hayden, 2015). From early childhood we discover causal relations through everyday play (Cook, Goodman, & Schulz, 2011; L. E. Schulz & Bonawitz, 2007), we explore based on how much we expect to learn (Bonawitz et al., 2011; Bonawitz, van Schijndel, Friel, & Schulz, 2012; Stahl & Feigenson, 2015), and we draw rational generalizations from limited observations (Gweon & Schulz, 2011; Xu & Garcia, 2008). As social creatures, however, we also rely on others to help us learn more than we could on our own (Bridgers, Gweon, Bretzke, & Ruggeri, 2018; Ruggeri & Lombrozo, 2015), seeking informants who are confident (Birch, Akmal, & Frampton, 2010; Brosseau-Liard & Poulin-Dubois, 2014), reliable (Poulin-Dubois, Brooker, & Polonia, 2011; Zmyj et al., 2010), and with a track record of being right (Koenig et al., 2004; Pasquini, Corriveau, Koenig, & Harris, 2007).

Despite the usefulness of social learning, identifying knowledgeable agents is a challenge in and of itself, requiring us to infer what others know based on how they behave. Nonetheless, adults routinely make quick and accurate guesses about others' knowledge from limited interactions. Imagine, for instance, asking a stranger on the street for directions to a nearby shop. If the stranger immediately told you where to go, you could reasonably assume that they know the place you're looking for, even if you couldn't immediately verify their answer. If instead, the stranger spent a painstaking amount of time consulting a map on their phone before telling you where to go, you could infer that the stranger hadn't heard about the shop you're looking for or didn't know its location.

Although these examples show that others' decisions to seek information can reveal what they know, many situations offer an even more nuanced glimpse into other people's minds. When the stranger consulted their map, the effort they invested in looking for information also reveals how much they cared about finding out the directions so that they could be helpful. Conversely, if the stranger gave you directions without consulting their phone, you might be more confident that their answer was accurate if they were leisurely sitting at a bench, phone in hand (and could have easily confirmed their directions before giving them), than if they were running late to a meeting, tried to check their phone but had poor reception, and pointed you in some direction before leaving abruptly (where their cost for confirming the directions would have been high).

These examples suggest that to infer knowledge, we consider not only whether others seek information, but also the costs associated with obtaining it, adjusting our inferences accordingly. Recent research suggests that even young children can make epistemic inferences from information-seeking behavior. Four- and five-year-olds judge that agents who can name animals without help are more likely to be knowledgeable relative to agents who accept help (Einav & Robinson, 2011), and they believe that agents who can state what's inside a container without checking are more likely to be knowledgeable, relative to agents that check before answer-

ing (Aboody, Huey, & Jara-Ettinger, 2018). While these studies show that children recognize the connection between information seeking and knowledge, it is unknown whether children also understand that agents' decisions about when to seek information are modulated by costs.

While epistemic inferences that incorporate costs might appear simple and intuitive to adults, they may not be obvious to children. On the one hand, even infants can integrate cost information to infer other people's mental states (S. Liu et al., 2017; Jara-Ettinger et al., 2016; Gergely & Csibra, 2003). However, prior work has restricted its focus to inferences about goals and desires: two mental-state representations that emerge very early in development (Gergely & Csibra, 2003; Woodward, 1998), where the logic of children's inferences is already structurally similar to that of adults (Baker, Saxe, & Tenenbaum, 2009; Lucas et al., 2014; Jern et al., 2017; Jara-Ettinger et al., 2020).

By contrast, representations of knowledge and belief develop later in childhood (Wellman et al., 2001; Wellman, 2014) and inferences about these mental states appear to be brittle and guided by simple heuristics. For instance, preschoolers do not recognize that ignorant agents will search randomly between two boxes (Chen et al., 2015; Friedman & Petrashek, 2009; Ruffman, 1996; Saxe, 2005)), they preferentially learn from familiar agents over accurate ones (Corriveau & Harris, 2009a), they overgeneralize knowledge onto a 'halo effect' (Brosseau-Liard & Birch, 2010), they fail to distinguish epistemic competence from non-epistemic competence (Fusaro, Corriveau, & Harris, 2011), they struggle to infer partial knowledge from partial goal-completion (Ronfard & Corriveau, 2016), and they incorrectly attribute expertise to agents who confidently answer questions that are impossible to answer correctly (Kominsky, Langthorne, & Keil, 2016). These heuristic-based inferences contrast with goal and desire inferences which, from infancy, are structured around an expectation of rational action that is sensitive to costs (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; S. Liu et al., 2017; S. Liu & Spelke, 2017).

Epistemic inferences that integrate others' information-seeking behavior with their costs would not only require children to break away from their typical use of heuristics in knowledge inferences, but also impose two difficult demands. First, such inferences require reasoning about how agents compare and balance quantities that are in fundamentally different metric spaces—information and cost. Second, they require children to represent how the cost of actions and the value of information vary across agents, depending both on agent-variable traits like physical competence or curiosity, and agent-variable mental states like goals or desires.

Here we test if children can infer others' epistemic states and desires by reasoning about rational trade-offs between agent-variable energy expenditure and information value. While substantial research has looked at children's own information-seeking behavior (Bonawitz et al., 2011, 2012; Cook et al., 2011; Ruggeri, Lombrozo, Griffiths, & Xu, 2016; E. Schulz, Wu, Ruggeri, & Meder, 2019; L. E. Schulz & Bonawitz, 2007; Stahl & Feigenson, 2015), less is known about children's epistemic inferences

from others' information-seeking behavior. Instead, the vast majority of research on children's action understanding has focused on reasoning about goals and preferences (e.g., Csibra, 2003; Jara-Ettinger et al., 2015; Lucas et al., 2014; Pesowski, Denison, & Friedman, 2016).

We present four experiments testing whether children make epistemic inferences through an expectation that agents rationally trade-off agent-variable costs and information value. We focus on a cost that even young children understand: physical effort (Jara-Ettinger, 2019; Leonard, Lee, & Schulz, 2017; S. Liu et al., 2017). In Experiment 1, we test if preschoolers believe that agents who refuse to seek (agent-variable) low-cost information are more likely to have already known it. In Experiment 2, we test if preschoolers believe that agents who seek (agent variable) high-cost information are more likely to have a strong desire for it. We focus on four- and five-year-olds because, although inferences based on action cost emerge early in infancy (Csibra, 2003; S. Liu et al., 2017), reasoning about agent-variable traits develops between ages five and eight (D. Liu, Gelman, & Wellman, 2007; Ruble & Dweck, 1995; Seiver, Gopnik, & Goodman, 2013). Moreover, our tasks require children to distinguish physical competence from epistemic competence, an ability which develops between ages three and five (Brosseau-Liard & Birch, 2010; Fusaro et al., 2011). We consequently also include two control experiments ruling out the possibility that children simply assume that stronger agents are more knowledgeable. Together, our experiments show that children expect agents to rationally trade-off information gain with costs, and that they use this expectation to infer others' knowledge based on agent-variable properties.

4.2 Sample Characteristics and Approach to Analyses

In line with current recommendations for statistical best practices, we take an estimation approach to data analysis rather than relying on null-hypothesis significance testing (Cohen, 1994; Cumming, 2014). We estimate effect sizes by bootstrapping our data and obtaining 95% confidence intervals; we take confidence intervals that do not cross chance as evidence of a reliable effect. Additionally, we use Bayesian data analyses to test whether our theoretical account can explain the full pattern of data obtained across all four experiments better than a simpler rule-based alternative.

We did not collect demographic information from participants, but report summary statistics based on their location of participation (obtained from census data). 56.4% of participants (n = 93) were recruited and tested in preschools in Los Angeles County. Computing demographic information by preschools' zip codes, median income for these areas is $78,812. On average 5.9% of adults in these areas were Black, 20.4% were Hispanic or Latino, 56.9% were White, 12.3% were Asian, 0.5% were Native American, 0.2% were Native Hawaiian or Pacific Islander, and 13.8% were two or more races, or marked "Other". 33.9% of participants (n = 56) were recruited and

tested at a museum in Boston, where on average 24% of attendees are Black, 17% are Hispanic or Latino, 47% are White, 9% are Asian, and 4% are two or more races. 29% of museum attendees visit on days when there is free or discounted admission. Finally, 9.7% of participants (n = 16) were recruited and tested at a museum in New Haven, where on average 19% of visitors are Black, 13% are Hispanic or Latino, 58% are White, 3% are Asian, 1% are Native American, and 6% are two or more races (Peabody Museum of Natural History, 2005). The median household income in New Haven is $42,222. All data were collected between May 2018 and January 2019.

4.3 Experiment 1

In Experiment 1, children watched a strong and a weak agent decline to lift a box to find out what was underneath. Participants were asked which of the two puppets already knew what was under the box. If children consider agent-variable tradeoffs between cost and information, they should infer that the stronger agent already knew what was inside. If, instead, children attend to information-seeking actions alone, they should perform at chance.

4.3.1 Methods

The procedure, predictions, and analysis plan were all pre-registered and are available at: `https://osf.io/27dkb/?view_only=7b7289c468e34e9f9cebf4cd78c85270`

Participants

48 four- and five-year-olds (mean age: 4.99 years, range: 4.20–5.88 years; n = 24 participants per age group) participated. Five additional participants were recruited but not included in the study (see Results). The pre-registered sample size for this and all following experiments was determined through a Monte Carlo power analysis (see Supplemental text; `https://osf.io/x26ha/`).

Stimuli

Stimuli consisted of two female puppets, two 5.75x5.75x5.25-inch gray boxes, and a small rubber duck. The boxes were closed at the top and open at the bottom, so items could be hidden underneath, and boxes could be lifted to reveal their contents. The first ("warm-up") box was empty and the second ("test") box had a rubber duck hidden underneath.

Procedure

Figure 4.1 shows the experimental procedure. The experimenter began by introducing the two boxes and the puppets, Adrienne and Sophie. The experimenter then

explained that "Adrienne is really strong, so it's easy for Adrienne to lift boxes." Adrienne then lifted the warm-up box swiftly on her first try, with no signs of exertion. The experimenter next explained that "Sophie is not strong, so it's very hard for Sophie to lift boxes." Sophie then struggled to lift the warm-up box, huffing with exertion and succeeding on the third try. Presentation order, and association between puppet (Adrienne or Sophie) and strength (weak or strong) were counterbalanced across participants.

The empty warm-up box was then removed and both puppets left the scene. The experimenter drew children's attention to the test box, saying, "And here we have a special box. This box is special because underneath, there's a rubber ducky!" She lifted the box to show participants the duck underneath, and then covered the duck again with the box. The experimenter then brought back the first puppet she had introduced, saying, "Let's give Adrienne a turn. Adrienne, there's something special under this box! Would you like to lift up the box, to find out what's underneath?" The puppet thought and said, "Hmm, no thanks!" The experimenter replied, "Ok!" and the puppet left the scene. Next, the experimenter brought back the second puppet and said "Let's give Sophie a turn. Sophie, there's something special under this box! Would you like to lift up the box, to find out what's underneath?" This puppet also thought and said, "Hmm, no thanks!" The experimenter replied "Ok!" and the puppet also left the scene.

Finally, the experimenter brought both puppets out and asked, "[Participant name], right before you came here today, one of my friends saw me put the rubber ducky under the box. So one of my friends already knew what was underneath the box. Can you tell me, which friend already knew what was underneath?"

The experimenter then asked participants to explain their choice (pre-registered as a variable not to be analyzed, but included for completeness), and then asked two inclusion questions: "Which one of my friends is really strong? And which one of my friends is not strong?"

4.3.2 Results

For the 88.7% of participants whose sessions were video or audio taped (n = 47/53), two coders who were not involved in data collection determined exclusions according to pre-registered criteria. The first coder, blind to participant answers, determined whether the experiment was run correctly. The second coder, blind to condition, coded participant answers. The experimenter took notes on any deviations from the procedure, and for participants who were not video or audio taped the first author determined exclusions by comparing these notes to the pre-registered inclusion criteria. Five participants were recruited but not included in the final sample because they incorrectly answered one or both of the inclusion questions (n = 2), because of experimenter error (n = 2), or because the participant took longer than 30 seconds to answer the test question (n = 1).

Out of the final 48 participants included in the study, 75% judged that the strong

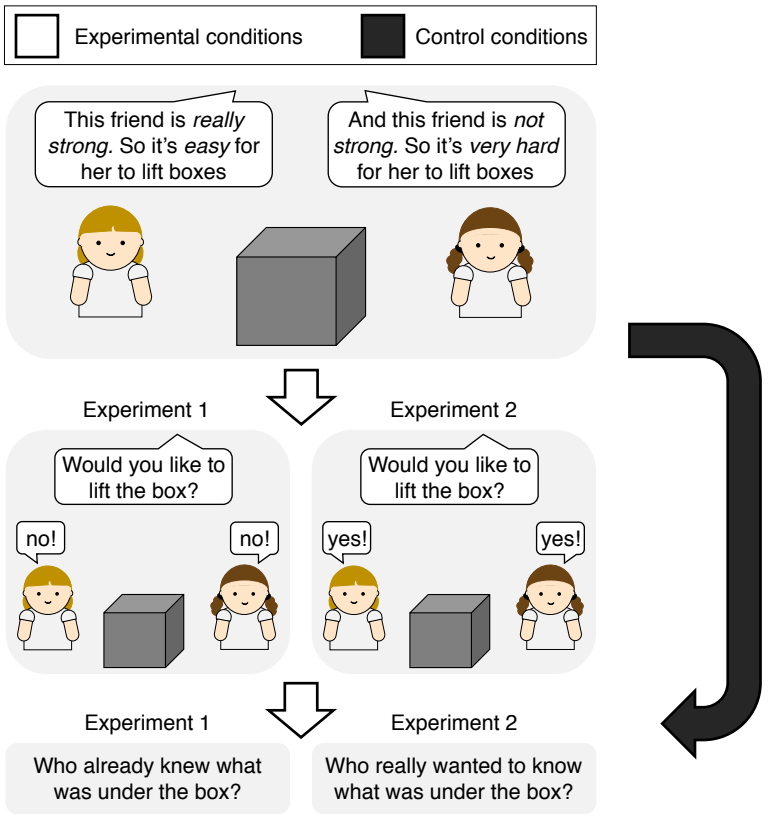

Figure 4.1: Schematic of Experiments 1–2 and their control conditions. After demonstrating their strength by lifting a first box, puppets were given the chance to lift a second box to find out what was underneath. Both agents refused to lift the box in Experiment 1, and they agreed to lift it in Experiment 2. The control experiments were identical to the main experiments, except that puppets were never given the chance to lift the second box. Instead, the experimenter proceeded directly to the test question.

agent already knew what was under the box (n = 36; 95% CI: 62.5 – 87.5; Figure 4.2). A logistic regression predicting performance as a function of age did not reveal any significant age difference (β = -0.65, p = .40), and performance within each age group was qualitatively similar: 79.2% of four-year-olds (n = 19 of 24) and 70.8% of five-year-olds (n = 17 of 24) judged that the strong agent already knew what was under the box. While young children often fail to produce relevant explanations in experimental contexts (Legare & Lombrozo, 2014; Walker, Lombrozo, Legare, & Gopnik, 2014; Walker, Bonawitz, & Lombrozo, 2017), many participants explained their answers by appealing to puppets' strength (see Supplemental Materials for explanations).

4.4 Experiment 2

Experiment 1 shows that when inferring knowledge from agents' exploratory choices, children consider the cost of seeking information. In Experiment 2 we test whether

children believe that agents who incur a higher cost to gain information must have a stronger epistemic desire. Children watched a strong and a weak puppet lift a box to find out what was underneath. Participants were asked which agent really wanted to know what was underneath. If children consider agents' costs when inferring their epistemic desires, they should infer that the weaker agent had a stronger desire to know. If, instead, they focus on outcome alone, they should perform at chance.

4.4.1 Methods

Participants

48 four- and five-year-olds (mean age: 5.04 years, range: 4.19–5.95 years; n = 24 participants per age group) were recruited. Eight additional participants were recruited but not included in the study (see Results).

Stimuli

Materials were the same as in Experiment 1.

Procedure

The procedure was nearly identical to Experiment 1, except that when given the opportunity to lift the test box, both puppets agreed, saying, "Hmm, okay!" The strong puppet lifted the test box with ease, and the weak puppet struggled but ultimately succeeded, as they had with the warm-up box (see Figure 4.1; presentation order, and association between puppet and strength was counterbalanced across participants).

After both puppets lifted the test box, the experimenter brought out the two puppets and asked "[Participant name], one of our friends really wanted to know what was under the box. Can you tell me, which friend really wanted to know?" Participants were then asked to explain their choice, followed by the same two inclusion questions from Experiment 1.

4.4.2 Results

Results were coded as in Experiment 1. 85.7% of participants were video or audio taped (n = 48/56). Eight participants were excluded from the final sample, because of interruptions or participant distraction (n = 3), experimenter error (n = 2), because the participant incorrectly answered one or both of the inclusion questions (n = 2), or because the participant took longer than 30 seconds to answer the test question (n = 1).

As predicted, children's pattern of responses flipped in Experiment 2. Out of the final 48 participants included in the study, 66.6% judged that the weak agent had the stronger epistemic desire (n = 32; 95% CI: 54.2 – 81.3). A logistic regression predicting performance as a function of age did not reveal any significant age difference

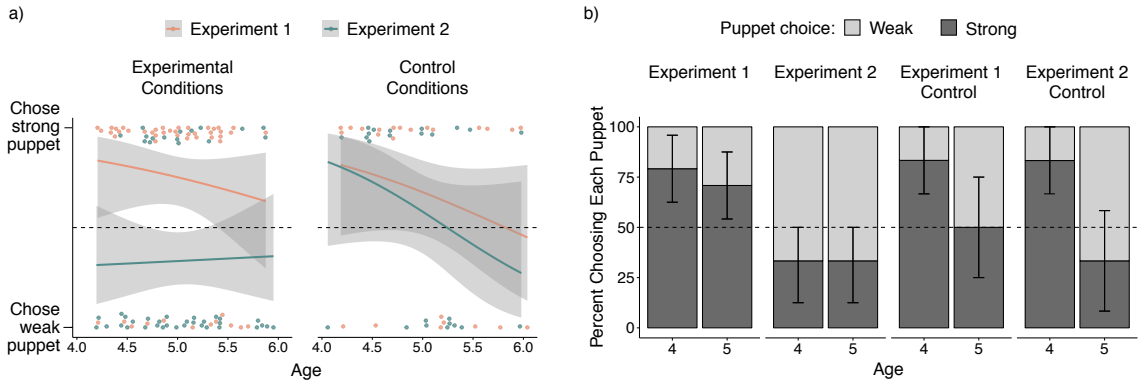

Figure 4.2: Results from all experiments. (a) Participant choices (strong puppet vs. weak puppet) as a function of age, along with logistic regressions fit to each data set. Points are jittered along the Y axis. The dotted line indicates chance performance. Gray bands show 95% CIs in the regression. (b) Participant choices were visualized by age group. Vertical bars show 95% bootstrapped confidence intervals.

(β = -0.11, p = .86), and performance within each age group was identical: 66.6% of four-year-olds (n = 16 of 24) and 66.6% of five-year-olds (n = 16 of 24) judged that the weak agent really wanted to know what was under the box. As in Experiment 1, participants who produced explanations often referred to puppets' strength (see Supplemental Materials for all explanations).

4.4.3 Experiments 1 and 2 Discussion

Experiments 1 and 2 suggest that children considered the cost associated with information gain when interpreting information-seeking behavior. By expecting agents to rationally trade-off agent-variable costs with agent-variable desires for information, children successfully inferred which of two agents was already knowledgeable when they both refused to obtain information (Experiment 1), and which of two agents most wanted knowledge when both agents chose to obtain information (Experiment 2).

Related research, however, has argued that children have a general representation of competence that combines strength, niceness, and knowledge (Brosseau-Liard & Birch, 2010; Fusaro et al., 2011). It is thus possible that children succeeded in Experiment 1 simply by assuming that stronger agents are more knowledgeable, and in Experiment 2 by assuming that weaker agents lack knowledge (and must therefore have stronger epistemic desires). Experiment 1 and 2 controls test for this possibility.

4.5 Experiment 1 and 2 Controls

The control conditions for Experiments 1 and 2 had identical procedures to the main experiments, with the difference that puppets were not given the opportunity to

lift the test box (Figure 4.1). Instead, the experimenter asked the (respective) test question immediately after the puppets had lifted the warm-up box (see Figure 4.1). If children's inferences in Experiments 1 and 2 were driven by a superficial assumption that stronger agents are more knowledgeable and that weaker agents are more curious, then the same pattern of results from Experiments 1 and 2 should appear in the control conditions.

4.5.1 Methods

Participants

Because we did not find any age difference in Experiment 1 and 2, we collapsed the two age groups and collected a single pre-registered sample of 24 four- to five-year-olds for each control experiment (Control Experiment 1: mean age: 5.01 years, range: 4.19–6.04 years; Control Experiment 2: mean age: 4.97 years, range: 4.06–5.98 years; n = 12 participants per age group per experiment). Eight additional participants were recruited but not included in the study (two in Control Experiment 1, and six in Control Experiment 2; see Results).

Stimuli

Materials were the same as in Experiments 1 and 2.

Procedure

The procedure for Experiment 1 control began in an identical way to Experiment 1. After the two puppets demonstrated their strength by lifting the warm-up box, the experimenter showed the participant (but not the puppets) that there was a rubber duck underneath the test box. However, instead of asking each puppet if they wanted to lift the box to find out what was underneath, the experimenter skipped straight to the test question ("which friend already knew what was underneath"), explanation prompt, and inclusion questions. Experiment 2 control was identical to Experiment 1 control with the difference that we matched the test question to the one from Experiment 2 ("which friend really wants to know?"). Note, however, that we switched the past-tense term "wanted" to present tense "wants", as puppets in this condition did not lift the test box (the action the past-tense "wanted" originally referred to).

4.5.2 Results

Results were coded in the same way as Experiments 1-2 (as pre-registered). In Control Experiment 1, 90% of participants were video or audio taped (n = 27/30), and in Control Experiment 2, 88.5% of participants were video or audio taped (n = 23/26). Eight additional participants were excluded and replaced, because they incorrectly

answered inclusion questions (Control Experiment 1, n = 2; Control Experiment 2, n = 2), or because they did not answer the test question within 30s (Control Experiment 1, n = 4).

In Control Experiment 1, 66.6% of participants judged that the strong agent already knew what was under the box, a proportion reliably higher than chance (n = 16 of 24, 95% CI: 50 – 87.5). A logistic regression predicting performance based on condition (control vs. experimental) revealed no significant effect of condition between Experiment 1 control and Experiment 1, (β = 0.41, p = 0.46).

In Control Experiment 2, only 41.7% of participants judged that the weak agent had the stronger epistemic desire, a proportion not reliably different from chance (n = 10 of 24; 95% CI: 20.8 – 62.5). A logistic regression predicting performance based on condition (control vs. experimental) revealed a significant effect of condition between Experiment 2 control and Experiment 2 (β = 1.03, p = 0.046). Participant explanations from both experiments are available in Supplemental Materials.

Combined, the results from the two control experiments suggest that a simple association between competence and knowledge cannot explain our full pattern of data. The strength-competence account predicts that children's performance in both control conditions should mirror performance in the experimental conditions, but children's responses significantly differed between Experiment 2 and its control.

The results above suggest that children in our main experiments flexibly adjusted their response based on the costs involved, while children in the control conditions did not. Consistent with this, a logistic regression predicting participant choice as a function of experimental condition (Experiment 1 vs. Experiment 2; not pre-registered) revealed a significant difference across conditions: participants were significantly less likely to select the strong agent in Experiment 2 (β = -1.79, p $<$ .001). By contrast, an equivalent (not pre-registered) regression predicting participant choice as a function of control condition found no significant difference across control conditions (β = -0.36, p = .55).

4.5.3 Control Experiments 1 and 2 Discussion

The results from our control conditions suggest that children's responses in our main studies were not driven by a simple strength-competence heuristic: If children assumed that strong agents are knowledgeable and weak agents desire knowledge (ignoring the costs that agents choose or refuse to incur), the pattern of results in the control conditions should have been identical to the pattern from the experimental conditions. Instead, children's responses in Experiment 2 significantly differed from their responses in the corresponding control condition.

4.6 Combined Bayesian Data Analysis

Our results suggest that children's epistemic inferences rely on their Theory of Mind, sensitive both to others' exploratory choices and their costs. However, these conclusions are based on analyses examining each experiment separately. To further evaluate both our theory and competing explanations, we tested how well each account could explain the entire pattern of data observed across all experiments (not pre-registered).

Formally, we considered three hypotheses: children select agents randomly (baseline model), children make epistemic inferences through a strength-competence heuristic (heuristic model), and children make epistemic inferences through their Theory of Mind (ToM model). Throughout, we use Bayes factors to compare theories, using standard terminology of Bayesian data analysis (Jeffreys, 1998).

To calculate the likelihood of the data given each theory, we took the product of four binomial distributions (one per data set), varying the probability of selecting each puppet according to each theory's predictions. The baseline model used a parameter of 0.5, expressing the prediction that participants had a 50% chance of selecting either puppet. The heuristic model used a parameter that tracked the chance a participant would use the strength-competence heuristic. For instance, a parameter of .75 would mean that each child had a 75% chance of selecting the strong agent in Experiment 1 and its control condition (judging that this agent was more knowledgeable) and a 75% chance of selecting the weak agent in Experiment 2 and its control condition (judging that this agent most desired knowledge). Finally, the ToM model used a parameter that indicated how children ought to perform when mental-state inference was possible, and predicted chance performance when cost information was absent. For instance, a parameter of .75 would mean that participants had a 75% chance of selecting the strong agent in Experiment 1 and the weak agent in Experiment 2, and a 50% chance of selecting either agent in the controls (as the ToM account makes no predictions in this case).

What factors might determine a participant's probability of success? In our pre-registered power analysis, we expected participants to succeed or fail based on a theory-independent feature: their attention. If this is the case, then the same proportion of participants should answer correctly no matter which theory they relied on (heuristic vs. ToM). Thus, in line with the pre-registered effect size that we expected, we began by setting the success probability to 75% in both the heuristic and ToM models. Using a uniform prior over hypotheses we found decisive evidence for the ToM model (BF = 3510.5 comparing ToM vs baseline; BF = 110.7 comparing ToM vs heuristic).

To test the robustness of our results, we also reproduced our analyses varying the expected probability of success of both models from 51% to 99% (in increments of 1%), and now additionally included the possibility that participants' probability of success differs based on the theory they relied on (e.g., it could be easier for participants to answer correctly if they relied on a simple heuristic rather than on their Theory of

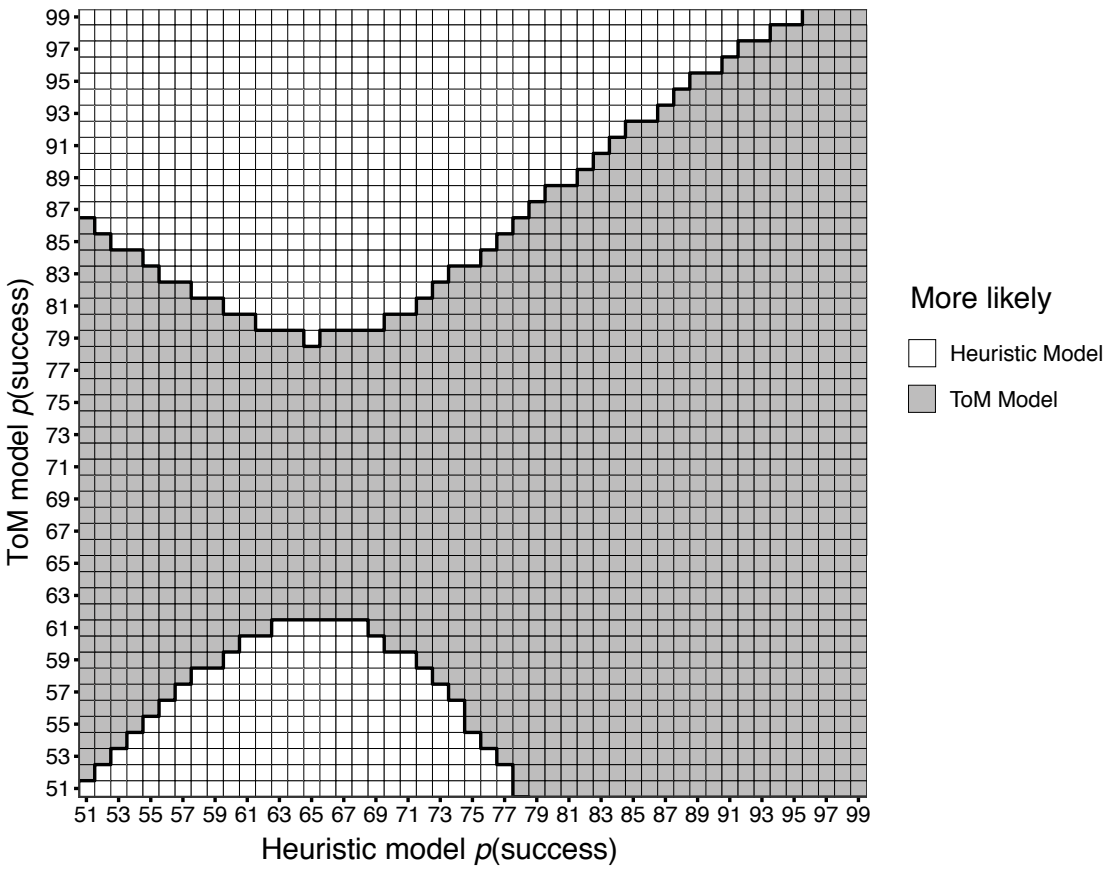

Figure 4.3: Each tile represents a comparison of the heuristic and theory of mind (ToM) models, given different parameter values for the expected probability of success. Dark gray tiles indicate cases where the ToM model is better able to explain the pattern of data (Bayes Factor > 1). White tiles indicate cases where the heuristic model is better able to explain the pattern of data (Bayes Factor < 1). For more information about the magnitude of the Bayes Factors, (see Supporting Information).

Mind). The ToM model outperformed the heuristic model in 67% of cases (n = 1,612 of 2,401), with a mean Bayes Factor of 2.99 x 1058, and a median Bayes Factor of 17.2 (see Figure 4.3; and Supplemental Materials for details).

Finally, we also conducted a full Bayesian model comparison that integrated uncertainty over the effect size. To achieve this, we placed a prior distribution over effect sizes centered at 75% success, and with a symmetrical shape (formally achieved by projecting a Beta distribution with parameters alpha=15 and beta=15 onto the .51-.99 range; see Supplemental Materials for details). Given this prior over effect sizes, and a uniform prior over theories (i.e., p(ToM)=p(heuristic)=0.5) we found strong evidence in favor of the ToM account (BF = 19.5). Additional robustness analyses showed that the qualitative conclusions are the same when the prior over effect sizes is relaxed (see Supplemental Materials).

Together, these results show that the ToM account was better at explaining our

data than the strength-competence heuristic. This held true no matter how we varied our parameters, whether we assumed that performance was theory-independent (i.e., the same across theories) or theory-dependent (i.e., different theories predicting different probabilities of success), and even when we integrated over these parameters using a prior over effect sizes.

4.7 General Discussion

As we navigate the social world, we must frequently infer what others believe and know from their actions. Such inferences can be far from straightforward: agents can produce the same action for different reasons, or pursue the same goal driven by different desires. Across four experiments, we showed that preschoolers infer agents' epistemic states and desires by considering how agents trade-off the agent-variable value of information with the agent-variable cost of obtaining it. In Experiment 1, four- and five-year-olds judged that an agent who declined to pursue low-cost information was more likely to be knowledgeable than an agent who declined to pursue the same information at a higher cost. In Experiment 2, children judged that an agent who incurred a higher cost to gain information must have wanted it more than an agent who incurred a low cost to obtain it. Two control experiments revealed that a superficial connection between strength and knowledge could not explain our results.

In all experiments, both agents always made identical decisions in identical situations. Thus, if children attempted to infer epistemic states on the basis of superficial observable cues, they should have performed at chance. Instead, our findings suggest that children attended to the psychological causes behind each agent's actions, taking into account their competence. Our findings add to a broader literature showing that, while children often make epistemic judgments on the basis of simple cues like accuracy or error (e.g., Koenig et al., 2004; Pasquini et al., 2007; Ruffman, 1996), they can nonetheless reason about the causes behind these cues when necessary (Aboody, Huey, & Jara-Ettinger, 2018; Einav & Robinson, 2011; Nurmsoo & Robinson, 2009a, 2009b).

Our work also sheds light on children's ability to represent compositional mental states. Research in Theory of Mind has typically focused on beliefs and desires as representations about the world (see Wellman, 2014, for a review). However, agents can also have beliefs about their desires (e.g., believing that they will like a new food) and desires about their beliefs (e.g., wanting to find out if their beliefs are true). To our knowledge, our work is the first to provide evidence that preschoolers can represent and infer desires about beliefs: In Experiment 2, children successfully identified the puppet that wanted to know. Along with research showing that preschoolers can also infer beliefs about desires (Jara-Ettinger et al., 2017), our results suggest that the ability to combine mental-state representations in a compositional manner emerges early in development.

Finally, our results converge with related work showing that inferences about desires are structured around an expectation that agents maximize utilities—the difference between costs agents incur and rewards they obtain (Gergely & Csibra, 2003; Jara-Ettinger et al., 2016; S. Liu et al., 2017). Our work extends these findings, showing that similar utility-based computations also enable children to infer others' epistemic desires and states. Together, this suggests that by age four, mental-state inference is grounded in a unified expectation that agents quantify, compare, and maximize their physical and epistemic utilities.

Our work opens several questions. First, our study focused on children's ability to infer epistemic states from information-seeking actions and their costs. If participants' judgments were guided by a causal understanding of how agents' competence and knowledge combine to produce action, children should be able to use information about any two of these factors to infer the third. That is, children should also be able to infer an agent's costs from their epistemic state and information-seeking actions; and predict whether an agent is likely to seek knowledge based on their costs and epistemic states. In contrast to a heuristics-based perspective, our account predicts that children should be able to derive all of these inferences given their causal utility-based naïve theory, and future work will test this possibility.

A related open question is whether children make such inferences spontaneously. In our experiments, we explicitly highlighted agents' actions and their costs, and prompted children to make epistemic inferences. But when children are not explicitly prompted to consider costs, they might be more likely to rely on quick and simple heuristics, or may not derive any epistemic inferences at all. Similarly, our tasks used constrained situations where only a few mental-state explanations were available; it is possible that children might appeal to other non-epistemic explanations in more naturalistic situations (e.g., assuming that an agent was doing something for fun rather than to gain information). As such, our work shows that children understand the role of costs in information-seeking, but leaves open the question of whether they make such epistemic inferences spontaneously.

Further, our studies manipulated the cost of action by varying agents' strength. If children's judgments were guided by an abstract representation of costs, then children should be able to solve equivalent tasks involving different sources of cost, such as dexterity, time, and mental effort. For instance, a willingness to incur a high cost to solve a puzzle box intuitively reveals a strong desire to know what's inside. However, these inferences should emerge only when children understand a domain well enough to grasp its cost structure (for instance, understanding how difficult it is to solve different kinds of puzzle boxes), which may take time to develop (e.g., S. Liu, Cushman, Gershman, Kool, & Spelke, 2019; E. Richardson & Keil, 2020, for children's understanding of mental effort).

More broadly, agents' decisions to seek or confirm information also depend on agent-variable traits. For instance, while an anxious person might continuously check for their wallet despite knowing it's there, a careless one might leave the house without

even thinking to do so. Effective epistemic inferences must therefore integrate richer agent-variable traits. Our work leaves open the question of whether children can infer these agent-variable traits and adjust their epistemic inferences accordingly.

In addition, our work did not explore the distinction between information that is intrinsically rewarding and information that serves as a means to an end. An ability to distinguish between the two is critical for inferring agents' desires (do they care about what they are learning?) and deciding how to react accordingly (should we tell them more, or focus on helping them achieve their ultimate goal?). Future work will explore whether children are sensitive not only to trade-offs between agents' costs and rewards, but also to trade-offs between the different kinds of rewards that can motivate agents to seek information.

Finally, our work leaves open questions about the developmental trajectory of the inferences we report here. In particular, we did not anticipate any age effects in our control conditions, and thus pre-registered a single sample of 24 four- and five-year-olds in each experiment. However, post-hoc analyses suggested that children's responses in the control conditions may differ by age (Figure 4.2b): four-year-olds appeared to prefer the strong agent in both control conditions (10 of 12 selecting this agent in each control, 95% CI: 66.6 – 100), while five-year-olds showed no reliable preference for either agent (6 of 12 children selecting the strong agent in Control Experiment 1, 95% CI: 25 – 75; and 4 of 12 in Control Experiment 2; 95% CI: 8.3 – 58.3). While exploratory, these results are consistent with prior research suggesting that younger preschoolers associate strength and competence (e.g., Fusaro et al., 2011), but suggest that this association can be easily overridden when more information about agents' costs is available. Note, however, that our control experiments were not powered to detect age effects and it is possible that these qualitative differences could have arisen due to chance. Future work will investigate this possibility.

4.7.1 Conclusion

During our preschool years, we invest so much time and effort into learning about the world. To learn most efficiently, we often rely on others to teach us what they know. Across four experiments, we find that preschoolers already appreciate the heterogeneity in what others know or want to know, engaging in nuanced mental-state reasoning to determine others' epistemic desires and states. This capacity may be at the heart of epistemic social behavior, not only guiding our decisions about whom to ask or trust, but also allowing us to determine who to help, how to teach, and even who should've known better.

4.7.2 Acknowledgements

We thank the families who participated in this research. We thank Madison Flowers and Lindsay Stoner for help with coding. This work was supported by a Google

Research Award. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

Chapter 5

Can preschoolers estimate an agent's probability of success under different degrees of knowledge?

This chapter is based upon Aboody, Denison & Jara-Ettinger (2021). Children consider the probability of random success when evaluating knowledge. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*

Abstract

To infer what others know, we must consider under what epistemic states their actions were both rational and probable. We test whether preschoolers can compare the probability of different actions (and outcomes) under different epistemic states—and use this to evaluate what others know. Specifically, we investigate whether four- to six-year-olds (n = 180) expect a task with a low probability of random success to better reveal an agent's epistemic state, as compared to a task where success is assured. By age six, children preferentially assigned the task with a low probability of random success to gauge an agent's epistemic state (Experiment 1) and inferred prior knowledge from successful completion of this task (Experiment 2). Across both experiments, four- and five-year-olds had no reliable preference, although children in all age groups were able to identify which task was harder. These results suggest that by the end of preschool, but not before, children integrate an understanding of probability with epistemic reasoning—realizing that a task where success is assured cannot reveal what others know, whereas a task with a low probability of random success might.

5.1 Introduction

To navigate the social world, it is important to understand what others believe and know. Because we can never see into others' minds, we often must infer their epistemic states from their behavior. As adults, such inferences are commonplace: if your friend is hurrying to work but skips a shortcut, you might infer they don't know about it. And if a new lab-mate immediately presses the right buzzer to unlock the lab door—ignoring three identical but irrelevant buttons—you might infer that someone has already shown them around.

Despite their ubiquity, such epistemic inferences are far from straightforward. While it may be convenient to infer ignorance from failure (e.g., skipping a shortcut) and knowledge from success (e.g., opening a door), the relationship between epistemic states and actions is not always deterministic. For instance, if all four buzzers actually unlocked the lab door (or the correct buzzer was prominently labeled), your lab-mate's success might not clearly reveal how much they already knew, since they would have succeeded no matter what. And if the buzzer just broke, even someone who knows how to get in might not be able to open the door. Thus, to infer what others know, we cannot consider only the outcome of their actions—we must also consider the reasons behind them.

Recent research suggests that even preschoolers do not rely solely on action outcomes to infer what others know (Aboody, Huey, & Jara-Ettinger, 2018; Aboody, Zhou, Flowers, & Jara-Ettinger, 2019; Einav & Robinson, 2011; Jara-Ettinger et al., 2017). For instance, four- and five-year-olds infer that an agent who refuses to pursue minimally costly information must have already known it, and an agent who decides to pursue high-cost information must have really wanted it (Aboody, Zhou, & Jara-Ettinger, 2021). This and related work (Jara-Ettinger et al., 2017) suggests that by the end of preschool, children rely on an expectation that agents maximize utilities (selecting actions that will yield the greatest rewards and incur the fewest costs) to infer others' epistemic states or motivations.

However, as the buzzer example illustrates, the relation between knowledge and action is mediated by the state of the world. To identify the reasons behind others' actions, we must consider whether these actions were truly diagnostic of their epistemic states. That is, to accurately infer what others know, children must be able to identify when an action is only likely given knowledge (e.g., picking the right buzzer when three don't work)—and understand that the exact same action could be ambiguous under different circumstances (e.g., if all four buzzers open the door).

It is possible that even young children consider others' actions in context, comparing the probability of agents' actions under different epistemic states to decide whether their behavior is more consistent with knowledge than ignorance. Children are sensitive to probability from infancy, distinguishing probable from improbable outcomes in the first years of life (Denison, Reed, & Xu, 2013; Gweon, Tenenbaum, & Schulz, 2010; Xu & Garcia, 2008). Additionally, recent research into children's epistemic inferences suggests that preschoolers may consider the probability of a random

success at least qualitatively: for instance, attributing knowledge to an agent who can predict an otherwise difficult-to-guess outcome, but not to an agent who observes the outcome and simply describes it (Aboody, Huey, & Jara-Ettinger, 2018).

However, it is also possible that young children do not consider the probability of an agent's actions under different epistemic states when inferring knowledge. Although preschoolers are extremely attentive to the outcome of others' actions (preferring to learn from and endorse the testimony of agents who were previously accurate; Corriveau & Harris, 2009b; Harris, 2012; Koenig et al., 2004; Koenig & Harris, 2005), they do not always consider those outcomes in context. For instance, preschoolers are not always sensitive to the reasons behind others' errors, trusting (Bridgers et al., 2016) or distrusting (Nurmsoo & Robinson, 2009b) agents without distinguishing whether their past errors were justified. Further, it is unclear whether children fully understand what it means to be ignorant: preschoolers do not expect ignorant agents to search randomly between potential locations of an object (Chen et al., 2015; Friedman & Petrashek, 2009; Ruffman, 1996; Saxe, 2005), and attribute greater expertise to those who confidently answer unknowable questions rather than those who correctly demur (Kominsky et al., 2016).

While a small number of studies have found that children are sensitive to probability when inferring desires (Diesendruck, Salzer, Kushnir, & Xu, 2015; Kushnir, Xu, & Wellman, 2010), developmental psychologists have long noticed asymmetries in children's understanding of mental states like goals or desires, and epistemic states like beliefs. From the first years of life, children are able to both represent and infer others' goals and desires (Gergely & Csibra, 2003; S. Liu et al., 2017; Woodward, 1998). However, it is not until age four or five that children can explicitly represent others' beliefs (for a review see Wellman et al., 2001). And even preschoolers who can represent epistemic states may not fully understand how to infer them: recent research suggests that an ability to infer epistemic states continues to develop between age four and six (Aboody, Huey, & Jara-Ettinger, 2018; Aboody et al., 2019; Wu & Schulz, 2018). Thus, it is unclear whether children can leverage their early-emerging understanding of probability to evaluate others' epistemic states, in addition to inferring their desires.

In the current experiments, we test whether four- to six-year-olds consider the probability of a chance success when evaluating others' epistemic states. Specifically, in Experiment 1 we test whether children understand that asking an agent to complete a "diagnostic" task (with only a 25% chance of random success) would better reveal their knowledge state, in contrast to an "undiagnostic" task (where success is assured). In Experiment 2, we test whether children are more likely to attribute prior knowledge to an agent who successfully completes the same "diagnostic" task (as compared to the "undiagnostic" task). We focus on four- to six-year-olds because children's belief reasoning (Wellman et al., 2001; Wellman, 2014; Wu & Schulz, 2018) and understanding of ignorance (Friedman & Petrashek, 2009; Ruffman, 1996) is still developing during the preschool years. Furthermore, recent work suggests that an

ability to use probability to predict others' emotional reactions is still developing between age four and six (Doan, Friedman, & Denison, 2018).

5.2 Approach to Analyses and General Methods

Consistent with recent recommendations for statistical best practices, we take an estimation approach to data analysis (Cohen, 1994; Cumming, 2014). We estimate effect sizes by bootstrapping our data and obtaining 95% confidence intervals, taking confidence intervals that do not cross chance as evidence of a reliable effect.

The procedures, predictions, sample size, exclusion criteria, and analysis plan for all experiments were pre-registered. All pre-registrations, stimuli, data, and analysis files are available in the OSF project page: `https://osf.io/e6b5h/?view_only=13f04c99e81c4220b79d1d9a5bec650b`. The pre-registered sample size for all experiments was determined through a Monte Carlo power analysis.

5.3 Experiment 1

5.3.1 Method

Participants

90 four- to six-year-olds (mean age: 5.51 years, range: 3.96-6.9 years) participated. Eleven additional participants were recruited but not included in the study (see Results). All participants completed the experiment via an online video-chat research platform.

Stimuli

Stimuli consisted of a Powerpoint presentation, featuring a cartoon character of a girl, four blue boxes lined up on a blue background, and four green boxes lined up on a green background. Five of the boxes (four on one side, and one on the other) had a yellow marble hidden underneath; three of the boxes were empty (see Figure 5.1).

Procedure

Figure 5.1 shows the experimental procedure. The experiment always began with eight boxes appearing on the screen. On the left were four blue boxes lined up on a blue background, and on the right were four green boxes lined up on a green background. The experimenter began by pointing out the boxes, saying, "Look! There are blue boxes on the blue side, and green boxes on the green side. Let's look under all of the boxes!" Starting on the blue side, the experimenter lifted each box one at a time to reveal its contents. Participants saw that every box on the blue side had a marble underneath (the "undiagnostic" side). The experimenter described each box's

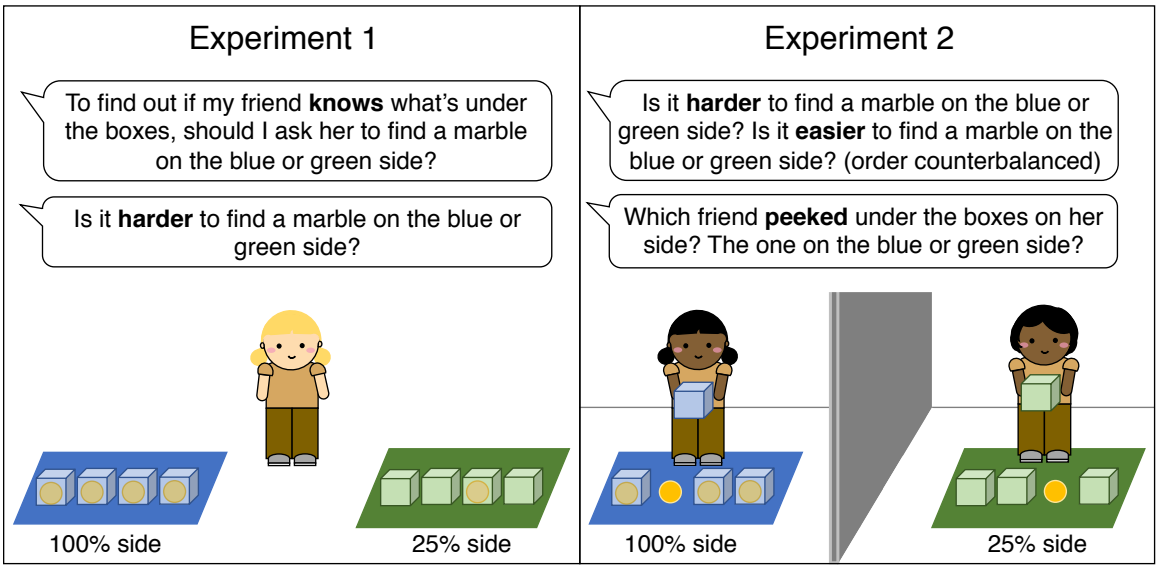

Figure 5.1: Procedure for all experiments. In all experiments, participants were first shown the contents of every box (not pictured). Next, in Experiment 1 the experimenter introduced a new friend, and asked participants to gauge her knowledge by asking her to find a marble on one of the sides. Participants were then asked to judge whether it was harder to find a marble on the blue side, or the green side. In Experiment 2, the experimenter introduced two new friends, and explained that one had already peeked under the boxes on her side, and that one had not looked under the boxes. The experimenter asked each in turn to find a marble on her side; both succeeded. Here, participants were first asked what was harder and what was easier, and next were asked to identify which one of the two friends had peeked and already knew what was under the boxes. Note: while we show each box's contents for clarity, in the experiment the boxes were opaque.

contents as they were revealed, saying, "Look, there's a marble under this box!" After lifting all of the boxes, the experimenter recapped by saying, "So, all of the boxes on the blue side have a marble underneath" (text italicized to mark words emphasized by the experimenter).

The experimenter then moved on to the green side, repeating the same procedure. Only one box on the green side had a yellow marble underneath (the marble was always under the third box), while the other boxes were empty (the "diagnostic" side). The experimenter described the box with the marble in the same way as before, and described the empty boxes by saying, "Look, there's nothing under this box." Finally, the experimenter recapped by saying, "So only one of the boxes on the green side has a marble underneath." The side with more marbles (blue vs. green) was counterbalanced across participants.

Next, a cartoon image of a child appeared in the middle of the screen, and the experimenter introduced the agent, saying, "Now, this is my friend. I want to find out if my friend knows what's under all of the boxes. Hmm. To figure out if my friend really knows what's under all of the boxes, let's ask her to show us a box that has a marble underneath. And we can see if she gets it right. We can ask our friend

to show us a marble on the blue side, or we can ask her to show us a marble on the green side." The experimenter continued on to the test questions, saying, "I need your help! I need to find out if my friend knows what's under all of the boxes. Should I ask her to find a marble on the blue side, or on the green side?" After participants responded, the experimenter asked them to explain their choice.

The experimenter then asked participants, "And which one is harder? Is it harder to find a marble on the blue side, or on the green side?" The experimenter again asked participants to explain their response, and finally asked the pre-registered inclusion questions, saying, "And can you remind me: which side had a lot of marbles? Blue or green? And which side only had one marble? Blue or green?" Note that although the blue side was always referenced first throughout the experiment, we counterbalanced whether this side was the correct option.

5.3.2 Results

For the 87.1% of participants whose sessions were video or audio taped (n = 88/101), two coders who were not involved in data collection determined exclusions according to pre-registered criteria. The first coder, blind to participant answers, determined whether the experiment was run correctly. The second coder, blind to condition, coded participant answers. The experimenter took notes on any deviations from the procedure, and for participants who were not video or audio taped the first author determined exclusions by comparing these notes to the pre-registered inclusion criteria. Eleven participants were recruited but not included in the final sample due to experimenter error (n = 3), technical difficulties (n = 2), because the participant did not provide codable answers to one or more questions (n = 2), failed an inclusion question (n = 1), was distracted (n = 1), did not wish to continue (n = 1), or due to interference (n = 1).

Out of the final 90 participants included in the study, only 57.8% of participants chose to evaluate the agent's knowledge by asking her to find a marble on the more diagnostic side (where only one of the four boxes had a marble underneath). This proportion is not reliably higher than chance (n = 52 of 90; 95% CI: 47.7 – 67.8). However, a logistic regression predicting performance as a function of age revealed a significant age difference ($\beta = 0.79$, p = .003), and performance within each age group qualitatively differed. Only 36.7% of four-year-olds (n = 11 of 30; 95% CI: 20 – 53.3) and 56.7% of five-year-olds (n = 17 of 30; 95% CI: 40 – 73.3) preferred to ask about the diagnostic side, whereas 80% of six-year-olds (n = 24 of 30; 95% CI: 66.7 – 96.7) did so (see Figure 5.2).

While only six-year-olds reliably wanted to ask the agent about the more diagnostic side, children of all ages understood that it was harder to find a marble on this side. 90% of participants (n = 81 of 90) correctly identified that it would be harder to find a marble on the diagnostic side, a proportion reliably higher than chance (95% CI: 84.1 – 96.6). A logistic regression predicting performance as a function of age did not reveal any significant age difference ($\beta = 0.35$, p = 0.39), and performance within

each age group was qualitatively similar. 83.3% of four-year-olds (n = 25 of 30; 95% CI: 70 – 96.7), 96.7% of five-year-olds (n = 29 of 30; 95% CI: 93.3 – 100), and 90% of six-year-olds (n = 27 of 30; 95% CI: 80 – 100) judged that it would be harder to find a marble on the diagnostic side (see Figure 5.2).

These results suggest that by age six, children realized that asking about the diagnostic side would better reveal an agent's epistemic state (as compared to the undiagnostic side). Younger children showed no reliable preference—although after the fact, they were able to judge that the diagnostic task was more difficult, suggesting they understood success on this side was less probable. Did younger participants simply fail to consider this information when deciding what to ask about, comparing the relative probability of an agent's success on each side only when explicitly prompted to do so? Or did participants realize that random success was less probable on the diagnostic side, but fail to use this information when trying to judge which side would best reveal what an agent knew?

To investigate, participants' explanations were coded post-hoc (not pre-registered) by the first author and another experimenter. Coders identified whether participants explicitly compared the diagnostic and undiagnostic sides, justifying their chosen side in reference to the other (e.g., "because it has more of the marbles"; "because it is more tricky"; "because there is only one marble under there")—or whether participants' answers simply described their chosen side, without explicit reference to the other (e.g., "because I saw four marbles under it"; "because it's easy"; "because there is a marble under there"). Uncodable explanations were designated as "other". Inter-rater reliability was high for all test questions ("what to ask" question: 86.7%; Cohen's $\kappa = 0.8$; $p < .001$; "what's harder" question: 97.8%; Cohen's $\kappa = 0.97$; $p < .001$). Disagreements were resolved by discussion.

Six-year-olds often compared the two sides, both when explaining which side they wanted to ask about (n = 20 of 30), and when explaining where it was harder to find a marble (n = 15 of 30). Almost all of these participants answered the relevant test question correctly (see Figure 5.2). Consistent with older children's performance, when explaining where it was harder to find a marble, every four- and five-year-old who referenced both sides (n = 21 of 60) correctly identified the "diagnostic" side as more difficult. However, when explaining which side they wanted to ask about, participants who referenced both sides (n = 15 of 60) performed poorly: a majority had actually wanted to see the agent find a marble on the incorrect "undiagnostic" side, where success was assured (9 of 15; see Figure 5.2).

These results demonstrate that even younger children sometimes explicitly contrasted the "diagnostic" and "undiagnostic" sides in their explanations. These participants always identified that the "diagnostic" side was harder—but often wanted to ask about the "undiagnostic" side to reveal the agent's epistemic state. This suggests that some four- and five-year-olds understood that both sides were not equally informative to ask about, but did not reliably consider or use this information when deciding how to evaluate the agent's knowledge state.

Figure 5.2: (a) Participant choices visualized by age group. The dotted line indicates predicted chance performance. Vertical bars show 95% bootstrapped confidence intervals. (b) Participant choices plotted continuously by age, along with a logistic regression fit to each dataset. Points are jittered along the Y axis (but not the X axis). (c) Of the participants who produced "comparison" explanations (referencing both the blue and green sides when explaining their responses), the proportion who answered the respective test question correctly. While most participants who produced "comparison" explanations were able to identify that the "diagnostic" task was harder, most younger participants who did so actually chose to ask the agent about the "undiagnostic" side.

5.3.3 Discussion

This experiment suggests that by age six (but not before), children realize that a task with a low probability of success will better reveal an agent's epistemic state, as compared to a task where success is assured. Four- and five-year-olds, however, had no reliable preference for the more diagnostic of the two tasks, although they were able to identify that this task would indeed be more difficult to complete (i.e., that random success is less probable). These results hint that even young children may be able to use probability to make objective judgments about features like difficulty—but not to gauge others' epistemic states.

Why might younger children have struggled in this task? While younger children may have genuinely struggled to integrate an understanding of probability with epistemic reasoning, there are at least two immediate alternative explanations. First, to interact with others in our environment, we need to both be able to predict how different mental states will lead agents to act, and infer others' mental states from their behavior. Some prior research suggests that, at least in certain cases, young children can identify what mental states caused a behavior earlier than they can predict how the same mental states might lead an agent to act (see Wellman, 2011). So it is possible that younger children struggled not with integrating epistemic states and probability, but simply with being asked to make an action prediction.

Second, it is also possible that younger children understood precisely how different epistemic states might affect an agent's probability of success on each side, but did not want to ask the agent to complete a difficult task. Children begin acting on

prosocial motivations early in life (Warneken & Tomasello, 2006); younger children might have decided to help the character in our task (rather than trying to select the task that would best reveal her knowledge state). Experiment 2 addresses both of these possibilities.

5.4 Experiment 2

5.4.1 Method

Participants

90 four- to six-year-olds (mean age: 5.5 years, range: 4.05-6.95 years) participated. Eleven additional participants were recruited but not included in the study (see Results). All participants completed the experiment via an online video-chat research platform.

Stimuli

Stimuli consisted of a Powerpoint presentation, featuring cartoon characters of two girls, four blue boxes lined up on a blue background, four green boxes lined up on a green background, and a wall separating the two. Five of the boxes (four on one side, and one on the other) had a yellow marble hidden underneath; three of the boxes were empty (see Figure 5.1).

Procedure

Figure 5.1 shows the experimental procedure. The procedure began nearly identically to that of Experiment 1, with the exception that after drawing participants' attention to the boxes, the experimenter also pointed out the wall in the middle of the screen, saying, "And look! There's a big wall in the middle. Do you see the wall? Great!" Without a wall separating the two sides, a knowledgeable agent could have chosen to position herself near the undiagnostic side simply because she knew it was easier to find a marble on this side—the addition of the wall minimized this concern.

After pointing out the wall, experiment always began with eight boxes appearing on the screen, and participants were introduced to the contents of the boxes in the same way as in Experiment 1 (the experimenter lifted each box one at a time). The only difference was that at the end, after lifting all of the boxes, the experimenter again repeated, "So, all of the boxes on the blue side have a marble underneath, and only one of the boxes on the green side has a marble underneath." We added this repetition as a conservative measure to ensure children remembered which side had more marbles (in light of younger children's surprising inability to identify that the "diagnostic" side was more helpful to ask about in Experiment 1). As before, the side with more marbles (blue vs. green) was counterbalanced across participants.

Next, the experimenter said, "Here I have two friends." An image of a cartoon child appeared on the left side of the screen, and the experimenter explained, "This is my friend Sally, on the blue side," (text italicized to mark words emphasized by the experimenter). Next, another child appeared on the right side of the screen, and the experimenter explained, "this is my friend Anne, on the green side." Continuing on, the experimenter said, "Now, right before you came here today, one of these friends peeked under the boxes. The other friend did not peek, and has never looked under the boxes. So one of these friends knows what's under the boxes on her side. And the other friend has no idea what's under the boxes on her side. I don't know if Sally peeked under all of the blue boxes, or if Anne peeked under all of the green boxes. Only one of my friends peeked. Hmm. To figure out which one of my friends peeked, let's ask each one a question."

The experimenter continued, "First, I'll ask my friend Sally. Let's figure out if Sally peeked under all the boxes on the blue side. Sally, can you find a marble on the blue side?" After asking this question, the experimenter showed participants that Sally had lifted the second box from the left, revealing a marble (see Figure 5.1). The experimenter said, "And look! Sally bent down and lifted this box. And she was right. So Sally found one of the boxes on the blue side that has a marble underneath!" Next, Sally put the box back, and the experimenter said, "Next, I'll ask my friend Anne. Let's figure out if Anne peeked under all the boxes on the green side. Anne, can you find a marble on the green side?" After asking this question, the experimenter showed participants that Anne had lifted the third box from the left, revealing a marble. The experimenter said, in the same tone as they used for Sally, "And look! Anne bent down and lifted this box. And she was right. So Anne found the only box on the green side that has a marble underneath!" Next, Anne put the box back.

The experimenter then reviewed the procedure, saying, "Remember how I told you that only one of these friends peeked under the boxes on her side? And the other one did not peek, and did not look under the boxes on her side? So only one of our friends knows what's under the boxes. Well, my friend Sally on the blue side found one of the boxes that has a marble underneath. And my friend Anne on the green side found the only box that has a marble underneath."

Next, the experimenter proceeded to the test questions. In contrast to Experiment 1, participants were first asked to both identify where it was harder to find a marble, and where it was easier (order counterbalanced). These changes were intended to prompt participants to consider the odds of success on both the blue and green side, and to do so before being asked to make an epistemic inference (with the intention to provide participants the best possible chance of making the appropriate inference).

The experimenter asked, "Can you tell me, which one is harder? Is it harder to find a marble on the blue side, or on the green side?" After repeating the participant's answer and eliciting an explanation, the experimenter asked, "And can you tell me, which one is easier? Is it easier to find a marble on the blue side, or on the green side?" The experimenter repeated the participant's answer and elicited an explanation. The

order of these two questions was counterbalanced. Finally, the experimenter asked children to make an epistemic inference, saying, "Now [participant name], I need your help! Can you tell me: who peeked? Was it Sally on the blue side, or Anne on the green side?" Again, the experimenter repeated the participant's answer and elicited an explanation. And as before, after the test questions the experimenter asked the pre-registered inclusion questions, saying, "And can you remind me: which side had a lot of marbles? Blue or green? And which side only had one marble? Blue or green?"

5.4.2 Results

For the 94.2% of participants whose sessions were video or audio taped (n = 93/101), two coders who were not involved in data collection determined exclusions according to pre-registered criteria, as in Experiment 1. Eleven participants were recruited but not included in the final sample due to distraction or inattention (n = 4), because the participant did not answer one or more questions (n = 2), failed an inclusion question (n = 1), due to experimenter error (n = 1), parental interference (n = 1), because the participant was too old (n = 1), or had already participated in the past (n = 1).

Out of the final 90 participants included in the study, 65.6% of participants judged that the agent who accurately completed the "diagnostic" task (finding the only marble on her side) had most likely peeked under the boxes. This proportion is reliably higher than chance (n = 59 of 90; 95% CI: 56.7 – 75.6). However, a logistic regression predicting performance as a function of age revealed a significant age difference (β = 0.61, p = .035), and performance within each age group qualitatively differed. Only 53% of four-year-olds (n = 16 of 30; 95% CI: 36.7 – 70) and 63% of five-year-olds (n = 19 of 30; 95% CI: 46.7 – 80) preferred to ask about the diagnostic side, whereas 80% of six-year-olds did so (n = 24 of 30; 95% CI: 66.7 – 96.7; see Figure 5.3).

While only six-year-olds reliably judged that the agent who successfully found a marble on the more diagnostic side (where only one box had a marble) was more likely knowledgeable than the agent who found a marble on the undiagnostic side (where every box contained a marble), children of all ages were able to identify which task was harder and which was easier. Of our participants, 83.3% (n = 75 of 90) correctly identified that it would be harder to find a marble on the diagnostic side. And 83.3% also correctly identified it would be easier to find a marble on the undiagnostic side. These proportions are reliably higher than chance (95% CI: 75.6 – 91.1). A logistic regression predicting performance as a function of age did not reveal any significant age difference for either test question (what's harder: β = 0.47, p = 0.19; what's easier: β = -0.06, p = 0.87). Consistent with this, performance within each age group was qualitatively similar: 73.3% of four-year-olds (n = 22 of 30; 95% CI: 60 – 90), 90% of five-year-olds (n = 27 of 30; 95% CI: 80 – 100), and 86.7% of six-year-olds (n = 26 of 30; 95% CI: 76.7 – 100) judged that it would be harder to find a marble on the diagnostic side (see Figure 5.3). Additionally, 83.3% of four-, five- and six-year-olds (n = 25 of 30, respectively; 95% CI: 70 – 96.7), judged that it would be easier to find a marble on the undiagnostic side (see Figure 5.3).

These results suggest that by age six, children realized it was improbable that an ignorant agent would find a marble on the "diagnostic" side—thus inferring that an agent who succeeded in doing so must have had prior knowledge.

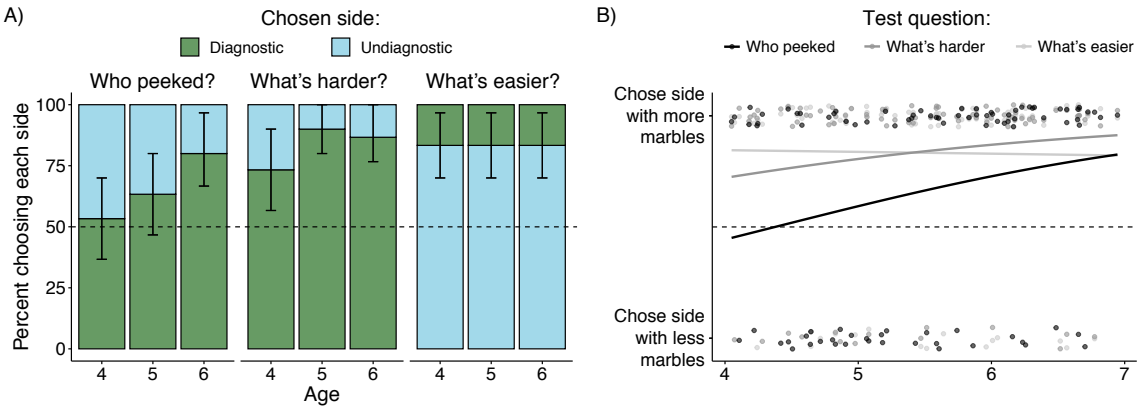

Figure 5.3: (a) Participant choices visualized by age group. The dotted line indicates predicted chance performance. Vertical bars show 95% bootstrapped confidence intervals. (b) Participant choices plotted continuously by age, along with a logistic regression fit to each dataset. Points are jittered along the Y axis (but not the X axis).

5.4.3 Discussion

This experiment suggests that by age six (but not before), children realize that an ignorant agent is unlikely to immediately succeed on a task with a low probability of success (as compared to a task where success is assured). Thus, six-year-olds inferred prior knowledge from an otherwise-improbable success. While four- and five-year-olds were able to identify that the "diagnostic" task would be harder to complete, and the "undiagnostic" task easier, they had no reliable preference when asked to infer which agent was knowledgeable.

These results demonstrate three things: First, while we considered the possibility that younger children struggled in Experiment 1 because explaining an agent's actions may be easier than predicting what they might do, this was not the case. Even after observing each agent's actions, children did not understand that success on the diagnostic task could most reasonably be explained by attributing prior knowledge. Second, we also considered the possibility that younger children in Experiment 1 simply wanted to help the agent succeed, and thus asked her to complete the undiagnostic task to ensure she would find a marble. This motivation cannot explain participants' failure in the current experiment, as participants were not offered a chance to help either agent (simply inferring each agent's epistemic state after already witnessing their successful action). Finally, we considered the possibility that participants did not compare the relative probability of success on each side unless explicitly prompted (although participants' explanations provided some evidence inconsistent with this possibility). In the current experiment, participants were asked to judge which task

was harder and which was easier before being asked to make an epistemic inference. Most children answered these questions correctly—and yet, younger children still did not judge that the agent who succeeded on the more difficult diagnostic task was more likely to be knowledgeable.

5.5 General Discussion

The capacity to teach or learn, help or hinder, and even punish or forgive, relies at least in part on an understanding of what others know. But the link between knowledge and action is highly variable, and is often mediated by the state of the world. For instance, if a friend arrives at their subway stop just as the train pulls into the station, you might infer they knew exactly when the train was due (and timed their walk accordingly). If they only caught the train because it was running late, you might be less charitable in your epistemic attribution. And if you find out that this train is actually late most of the time, then you might again suspect that your friend knows exactly what they're doing. Successful knowledge inferences thus require us to consider others' actions in context, comparing how different degrees of knowledge might lead them to act in any given situation.

We show that by the end of preschool, children understand how epistemic states and the environment interact, deciding what others know by comparing the likelihood of their actions under different degrees of knowledge. Specifically, when asked to gauge an agent's knowledge state in Experiment 1, six-year-olds wanted to see her complete a "diagnostic" task with a low probability of random success (rather than an "undiagnostic" task, where success was assured). And upon observing an agent successfully complete each task on the first try in Experiment 2, six-year-olds inferred that the agent who completed the "diagnostic" task likely had prior knowledge (as compared to the agent who completed the assured "undiagnostic" task). Across both experiments, four- and five-year-olds' judgments did not reliably differ from chance, although younger children were still able to identify which task was more difficult (and also which was easier, in Experiment 2).

Thus, our results suggest that before age six, children may not integrate an understanding of probability with epistemic reasoning to predict how different epistemic states may lead agents to act (or infer these epistemic states from agents' actions). However, by age six (although perhaps not before), children understand that even an ignorant agent will succeed in a case where success is assured—but that only a knowledgeable agent is likely to succeed in a case where random success is improbable.

At first glance, four- and five-year-olds' difficulties may be surprising. After all, prior research shows that children are sensitive to probability from infancy (Denison et al., 2013; Gweon et al., 2010; Xu & Garcia, 2008), and infer others' intentions and desires by considering the probability of their action outcomes from the first years of life (Diesendruck et al., 2015; Kushnir et al., 2010). However, our results are consistent with a body of related work which suggests that an ability to integrate

probability and belief to predict others' emotions is still developing between age five and seven (Doan et al., 2018; MacLaren & Olson, 1993; Ruffman & Keenan, 1996; but see Scott, 2017). For instance, six-year-olds can accurately predict whether an agent will be surprised when asked to reason about the objective probability of an outcome—but not when prompted to consider an agent's belief over this outcome (Doan et al., 2018).

The apparent divide in children's use of probability in mental-state reasoning is also consistent with the broader development of children's Theory of Mind. While children can infer others' goals, intentions, and desires from the first years of life (Gergely & Csibra, 2003; Meltzoff, 1995; Woodward, 1998), it is not until age four or five that children can reliably and explicitly represent others' false beliefs (Wellman et al., 2001). Younger preschoolers may thus have less experience reasoning about epistemic states, as compared to mental states like desires, and it may take them longer to fully understand how epistemic states and action relate. This possibility is consistent with research finding that a full understanding of epistemic states may continue to develop throughout the late preschool years (e.g., Aboody et al., 2019; Wu & Schulz, 2018). For instance, it is not until age six that children even begin to appreciate that ignorant agents will search randomly (Chen et al., 2015; Friedman & Petrashek, 2009; Ruffman, 1996). If four- and five-year-olds in our task did not expect that an ignorant agent would search randomly for a marble, this could explain why younger children did not prefer to ask about—and did not infer knowledge from success on—the diagnostic task. This possibility highlights the need for further research to investigate not only how children represent beliefs, but also how young children reason about and infer epistemic states like knowledge and ignorance (see also Phillips et al., 2021).

Finally, by internally replicating our age trajectory across two experiments, our results provide further evidence that the nuances of any single task may not explain our full pattern of data. Specifically, Experiment 1 opened several possible alternate explanations for younger children's failures: 1) That younger children might make epistemic inferences more readily than action predictions (e.g., see Wellman, 2011); 2) That a drive to be prosocial might have influenced some children's judgments, causing them to try to help the agent find a marble by asking her about the side where her success was assured; and 3) That participants may have been able to identify which task was harder, but only when explicitly prompted to do so. These alternatives cannot explain younger children's failures in Experiment 2: participants were asked to make epistemic inferences, could not help the agent (having already witnessed both agents succeed), and were asked to first explicitly identify which task was easier and which was harder before judging who was knowledgeable. By ruling out several reasonable alternative explanations for our results, our work again points to the possibility that younger children failed because they truly struggled to reason about the link between epistemic states and action; future work should continue to investigate how children make action predictions from epistemic state information.

Our work also leaves open at least two further questions. First, we found that by age six, children both expect a task with a lower probability of random success to better reveal an agent's knowledge state, and indeed infer prior knowledge from immediate success on this task (as compared to a task where success is assured). While this shows that older children at least realized that an ignorant agent would likely struggle on the diagnostic task, this may not be because they truly considered the probability that an ignorant agent would succeed or fail. For instance, older children could have simply preferred to ask about the only side where failure was possible (and inferred knowledge from a success on this task) without considering precisely how probable success or failure was. Or participants could have relied on an even simpler heuristic, assuming that ignorant agents make mistakes (e.g., Ruffman, 1996), and thus selecting the only side where error was possible. Although six-year-olds' explanations suggest that many participants did explicitly compare the agent's probability of success on each task (n = 20 of 30 in Experiment 1), future work will test for this possibility (for instance, by replicating Experiment 2 in a fully probabilistic task, contrasting a "less diagnostic" side with three marbles to a "more diagnostic" side featuring only one).

Second, it is worth noting that even as adults, we do not always infer knowledge from an otherwise unlikely success (most of us would agree that a person who picked the winning lottery number didn't know they would win). Conversely, even adults can struggle to correctly estimate the probability of a successful action under ignorance (for instance, attributing otherworldly knowledge when an astrological prediction or palm reading happens to describe their life). And sometimes we rely on trait attributions (rather than epistemic or world states) to explain others' successes or failures; for instance, describing a person as "unlucky" if they have experienced a string of otherwise unexpectedly poor outcomes. It is unclear how young children reason about such edge cases, especially in situations where even adults can struggle.

5.5.1 Conclusion

To infer the cause of a failed action, figure out what to teach, or even decide who knew better, we must understand others' epistemic states. In the current work, we find that by age six, children understand that the state of the world mediates the relation between knowledge and action—using this to decide under what conditions an action or outcome truly reveals knowledge. These results highlight the complexity of everyday epistemic judgments, and the need for further research into children's understanding of the relation between knowledge, ignorance, belief, and action.

5.5.2 Acknowledgments

We thank Colin Jacobs, Sofia Rubio, Eden Senay, and Hudson Patterson for assistance with data collection. We thank Ilayda Orhan and Rodney Tompkins for assistance with coding. We thank Katie Vasquez for her collaborative approach to participant

recruitment. This work was supported by Yale's Franke Program in Science and the Humanities, via a Franke Interdisciplinary Graduate Award to RA.

Chapter 6

Can preschoolers estimate the difference between an agent's probability of success, given different knowledge states?

This chapter is based upon Aboody, Flowers, Zhou & Jara-Ettinger (2019). Ignorance = doing what is reasonable: Children expect ignorant agents to act based on prior knowledge. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*

Abstract

When deciding how to act in new situations, we expect agents to draw on relevant prior experiences. This expectation underlies many of our mental-state inferences, allowing us to infer agents' prior knowledge from their current actions. Do children share this expectation, and use it to infer others' epistemic states? In Experiment 1, we find that five- and six-year-olds (but not four-year-olds) attribute additional knowledge to agents whose prior experiences cannot explain their success. In Experiment 2, we find that six-year-olds (but not younger children) also attribute greater knowledge to agents whose prior experience cannot explain their failure. We show that by age five or six, children expect ignorant agents' beliefs (and therefore their actions) to be guided by their prior knowledge. This work adds to a growing body of research suggesting that, while infants can represent mental states, the ability to infer mental states continues to develop throughout early childhood.

6.1 Introduction

To discuss someone's ambitions, frustrations, or disappointments is to talk about a mind that works much like our own, except that we cannot see it or know what it knows. Yet, we make surprisingly accurate inferences about what others think or want, just by watching how they act. For example, if your friend gives you her keys but later rummages in her bag upon reaching the car, you might infer that she forgot you have them. If she doesn't slow for a pedestrian at a crosswalk, you'd probably assume she didn't see them. And if she suddenly takes a detour, you might suspect she knows something you don't (perhaps the usual route is under construction).

The ability to infer other people's thoughts and desires from their behavior involves building a working model of how their mental states relate to their actions. The foundations of this capacity, called a Theory of Mind (Dennett, 1987; Gopnik & Wellman, 1992), are in place and at work early in infancy (S. Liu et al., 2017; Woodward, 1998) but continue to mature throughout early childhood (Wellman et al., 2001), and well into adolescence (H. Richardson et al., 2018).

Within Theory of Mind, our ability to reason about other people's beliefs—what they know, what they don't, and what they think they know—is particularly slow to develop. While infants can represent other people's beliefs (Onishi & Baillargeon, 2005), knowledge (Surian, Caldi, & Sperber, 2007), and ignorance (O'Neill, 1996), children do not use these representations explicitly until several years later (Bartsch & Wellman, 1995; Wellman et al., 2001).

As adults, we understand that other people's past experiences shape their current beliefs, and that these beliefs guide their actions. If, for example, your friend starts their car by inserting and turning a key, you can reasonably predict they will try the same the first time they drive yours. And you'd expect this even if you know your car works differently (for example, starting when a button is pushed in proximity to the key fob).

This expectation not only allows us to predict how others will act: it also allows us to infer what they know by observing how they act. In the example above, if your friend defied your expectations by immediately locating the button that starts your car, you might wonder if they had some prior experience you didn't know about (perhaps they've driven other cars like yours before). Such reasoning may seem intuitive, but how exactly do we predict what actions agents are likely to take in new situations? Prior research suggests that adults solve this problem by integrating over agents' uncertainty (Baker et al., 2017). For instance, when we reason about an agent who does not know whether a car starts via a key or a button, we consider what they would do in each situation, and we expect them to choose a plan weighted by their confidence.

While effective, these types of inferences are computationally complex. They require considering multiple possible worlds (at least implicitly), and deciding what an agent would do in each. Perhaps unsurprisingly, children's expectations for how ignorant agents are likely to act appear to rely on simpler strategies. Children some-

times equate being ignorant with getting things wrong (Ruffman, 1996; Saxe, 2005); although, in other contexts, their intuitions reverse (Friedman & Petrashek, 2009; German & Leslie, 2001).

While expecting ignorant agents to fail may support accurate inferences and useful predictions, such strategies are limited. Even ignorant agents can make reasonable guesses based on past experience. For instance, even if you've only used PC's, you probably have some idea of what you'd try if you had to turn on a Mac. And ignorant agents can always get lucky, succeeding by chance.

Do children understand how previous experiences affect agents' future actions? And do they leverage this expectation to infer what an agent knows based on what she does? In the current work, we investigate these questions with four- to six-year-olds. The ability to explicitly and flexibly represent beliefs emerges in the mid-preschool years (e.g., Rubio-Fernández, 2019; Wellman et al., 2001. Therefore, if children have expectations about the relation between ignorance and action, we might expect them to emerge in this age range.

In two experiments, participants watched two puppets learn how to activate a novel toy. Each puppet later attempted to activate a different (but outwardly identical) toy. One agent's actions were consistent with their prior experience, while the other agent's actions were inconsistent with their prior experience. In Experiment 1, both agents succeeded in activating a toy. If children expect agents to act based on their prior knowledge, they should judge that the inconsistent agent (whose actions cannot be explained by their experience with the initial toy) must have had additional knowledge. We find that five- and six-year-olds (but not four-year-olds) attribute additional prior knowledge to this agent.

To control for the possibility that children attribute knowledge to agents who teach them something new, in Experiment 2, children learned how a toy worked, and then watched two agents fail to activate this toy. Children again judged that the inconsistent agent (whose action couldn't be explained by his experience with the initial toy) had greater additional knowledge. These results suggest that by age five, children expect ignorant agents to act according to their prior knowledge, and further, that children leverage this expectation to infer what others know from what they do. All experiments' procedures, predictions, exclusion criteria, and analyses were pre-registered.

6.2 Experiment 1

In Experiment 1, children watched two puppets learn how to activate a novel toy. Next, each puppet was given the chance to activate a different toy (always outwardly identical to the original). One puppet stated that his chosen toy worked the same as the original, and pressed the same button he had seen activate the original toy. The other puppet stated that his chosen toy worked differently to the original, and pressed a different button. Both puppets succeeded in activating their chosen toy.

Children were then asked which of the two agents already knew how the toys worked.

If children expect ignorant agents to behave in accordance with their prior beliefs, then they should judge that the agent who acted inconsistently with their prior experience is more likely to be knowledgeable. But if children attribute epistemic states by relying on a rule of thumb (e.g., expecting ignorant agents to be wrong), or have no representation of what it means to be ignorant, then children should have no preference for either agent.

6.2.1 Method

Participants

72 four-, five-, and six-year-olds (mean age: 5.46 years, range: 4.05 – 6.99 years; n = 24 participants per age group) were recruited at a local children's museum. 22 participants were excluded from the analyses and replaced because: they did not pass the pre-registered inclusion questions (n = 9), due to experimenter error (n = 5), interruptions from other children (n = 3), because the participant did not answer the test question within 30s (n = 2), distraction (n = 1), interference with the procedure (n = 1), or due to developmental delays (n = 1).

Stimuli

Stimuli consisted of two male puppets, and three novel toys. These toys were externally identical machines, each covered in black construction paper and measuring approximately 5 x 3 x 2.75 in. Toys had three buttons on top: a red button in the middle, and two black buttons flanking the red one (see Figure 6.1).

Although they all looked the same, the toys worked in different ways. The first toy (called the "training" toy) activated and played music only when the central red button was pressed. Of the remaining toys, the "consistent" one worked the same way. However, the "inconsistent" toy worked differently: only pressing the black button to the participant's far left made it activate. For clarity, we refer to this button as the "correct" black button, and the other as the "incorrect" black button (since it did not activate the toy).

Procedure

First, participants were familiarized with the training toy (which turned on when the central red button was pressed). Participants learned that the red button made the toy go, but that the black buttons did nothing. They were then given a chance to press all of the buttons themselves. Next, participants were introduced to two puppets. The experimenter explained that she was going to show the puppets how the toy worked, and told the puppets that while the red button made the toy go, the black buttons did not do anything. Upon the experimenter's request, the puppets pressed the red button together.

Next, the remaining toys were placed on the table (one on either side of the training toy). The experimenter explained that one of the puppets had snuck out from under the table and played with all the toys, and discovered which buttons made the toys play music. The other puppet had stayed underneath the table, and hadn't seen anything. The child's task was to help figure out which puppet had snuck out and played with all the toys.

Each puppet was questioned individually, while the other agent was placed under the table. During his turn, each puppet was asked: "Can you show us how to make one of these toys go?" To make the relation between agents' actions and their experience with the initial toy more explicit, each agent explained himself as he acted. One puppet chose the consistent toy, saying, "Hmm. Well, the red button made this [original] toy go, so the red button makes this toy go too," pointing to the two relevant buttons as he spoke. Finally he pressed the red button, successfully activating the toy. The other puppet chose the inconsistent toy, saying, "Hmm. Well, the red button made this [original] toy go, but this black button makes this toy go," pointing to the two relevant buttons as he spoke. Finally he pressed the correct black button, successfully activating the toy.

After each puppet demonstrated one of the toys, the experimenter asked the test question: "[Child name], remember how I told you at the beginning of the game that only one of my friends snuck out from underneath the table, and played with all the toys? Can you tell me, which one of my friends snuck out and played with all the toys?" Participants were then asked to explain their answer. The memory check questions (pre-registered as inclusion questions) were asked last, with subjects asked to match each puppet to the toy he had demonstrated: "[Child name], can you remind me, which friend showed us how to make this toy go [both puppets point to a toy]? And which friend showed us how to make this toy go [both puppets point to the other toy]?"

Puppets always demonstrated the toy they were standing closest to. This was to avoid pragmatic concerns that could arise if puppets undertook a cost to demonstrate a particular toy. Therefore, the puppet on the experimenter's left hand demonstrated the leftmost toy, and vice versa. The identity of the puppet whose turn was first, and the toy this agent acted on was always counterbalanced. Additionally, the side each puppet was presented on (left/right) was randomized.

6.2.2 Results and Discussion

Two coders who were not involved in data collection determined exclusions. The first coder determined whether the experiment had been run correctly, blind to children's final answers. The second coder coded only children's answers, unaware of each puppet's role (that is, whether he demonstrated the consistent or inconsistent toy). 22 participants were excluded and replaced (see Participants).

Overall, of 58.3% of children judged that the agent who pressed the black button (and acted inconsistently with his prior experience) was more likely to have had

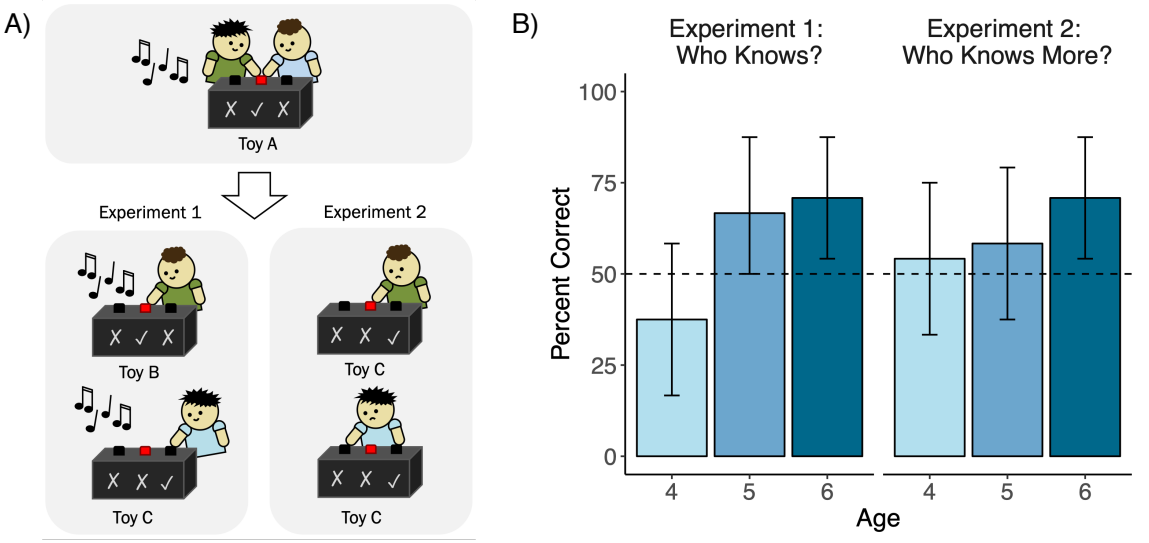

Figure 6.1: A) Procedure of both experiments. In Experiment 1 both puppets succeeded in activating the toy. In Experiment 2, both failed. Crucially, one agent's actions were always consistent with his prior experience (pressing the red button); the other agent's were not (he pressed one of the black buttons). B) Results from both experiments. The error bars are bootstrapped 95% confidence intervals, and the dotted line indicates chance performance (50%). In Experiment 1, five- and six-year-olds judged that an agent whose success could not be explained by his prior experience had additional knowledge. In Experiment 2, six-year-olds judged that an agent whose failure could not be explained by his prior experience had additional (albeit incomplete) knowledge.

additional knowledge. This proportion is not reliably different from chance (42 of 72; 95% CI: 47.2 – 69.4). However, a logistic regression predicting performance based on age revealed a significant age difference ($\beta = 0.87$, p = .006). While only 37.5% of four-year-olds judged that the agent who activated the inconsistent toy had prior knowledge (9 of 24; 95% CI: 16.67 - 58.33), 66.6% of five-year-olds (16 of 24; 95% CI: 50 – 87.5) and 70.8% of six-year-olds (17 of 24; 95% CI: 54.17 - 87.5) selected this agent. And consistent with five- and six-year-olds' success, a logistic regression predicting performance based on age also predicts that children will be more likely to answer the test question correctly (as opposed to incorrectly) by 5.04 years of age.

These results suggest that children do not simply expect ignorant agents to act successfully or unsuccessfully. Rather, by age five, children seem to expect ignorant agents to act reasonably, applying their prior knowledge in novel situations. This is consistent with prior findings that children do not think ignorance means having a false belief (Friedman & Petrashek, 2009; Jara-Ettinger et al., 2017). If children assumed that ignorant agents should fail due to a false belief, then participants should have judged that both agents were equally knowledgeable (since both were successful). Our results suggest that by age five, children make principled belief inferences from agents' behavior. Specifically, children expect both knowledgeable and ignorant agents to act consistently with their prior knowledge, and they use these expectations

to infer what other people know.

Note, however, that children were only ever taught how the training toy worked. If children (reasonably) assumed all the toys worked in the same way, they may have been surprised to see a puppet activate the inconsistent toy. Perhaps children attributed greater knowledge to this agent not because his actions were inconsistent with his prior knowledge, but because the actions (and their outcome) were inconsistent with children's own beliefs. In other words, children might simply attribute knowledge to agents who teach them something new, or show them something unexpected. We test this possibility in Experiment 2.

6.3 Experiment 2

Participants in Experiment 1 learned only how the first (training) toy worked. If participants attributed greater knowledge to the inconsistent actor because he taught them something new or unexpected, teaching children how all the toys work should cause performance to fall to chance because, now, neither agent can provide any novel information.

To address this, Experiment 2 differs in three substantial ways. First, we taught participants how all the toys worked. To reduce concerns about memory load, we used only two machines in this task: the training toy, and the inconsistent toy. Second, when trying to activate the novel toy, both puppets failed. One puppet pressed the red button (consistent with his prior experience), and one pressed the incorrect black button (inconsistent with his prior experience). Finally, we emphasized throughout that one of the puppets knew more, but not all, about the toy, making it plausible that both puppets could fail. Together, these changes allow us to test whether children attribute greater prior knowledge to agents whose actions are not explained by their prior experience, even when the agent fails to achieve their goal.

6.3.1 Method

Participants

72 four-, five-, and six-year-olds (mean age: 5.56 years, range: 3.99 – 6.92 years; n = 24 participants per age group) were recruited at a local children's museum. 26 participants were excluded from analyses and replaced because: they did not pass the pre-registered inclusion questions (n = 13), due to experimenter error (n = 5), interruptions or interference with the procedure (n = 3), because the participant did not answer the test question within 30s (n = 3), because the participant had already participated in the past (n = 1), or due to developmental delays (n = 1).

Stimuli

Materials were identical to those of Experiment 1, except that now only two machines were used: the training toy, and the inconsistent toy.

Procedure

Experiment 2 began identically to Experiment 1. Participants and then puppets were familiarized with the training toy. Next, after placing the puppets underneath the table, the experimenter produced the additional (inconsistent) toy. In contrast to Experiment 1, the experimenter told participants that this toy was "a little bit different." She explained that the red button did not activate this toy, and that only one of the black buttons (the correct black button) made the toy play music. She demonstrated all of the buttons, and then allowed the participant to press each button. Thus, participants were explicitly taught how the toys worked, and experienced for themselves that the toys worked differently.

Next, both puppets returned. The experimenter explained that one of the puppets had seen the toy before, and knew a little bit about it. And she explained that the other puppet had never seen the toy before. The experimenter noted that one of the puppets knew more about the toy, but she didn't know which one. The participant's task was to help the experimenter identify which puppet knew more about the toy.

Each puppet was asked to make the toy go in turn. During each puppet's turn, the other agent was placed underneath the table. One puppet's actions were consistent with his prior knowledge, saying, "Hmm. Well, the red button made this [original] toy go, so the red button makes this toy go too," pointing to the two relevant buttons as he spoke. He pressed the red button. The button did not activate the toy, and the puppet exclaimed "oh!" in surprise when nothing happened. The other puppet's actions were inconsistent with his prior knowledge, saying, "Hmm. Well, the red button made this [original] toy go, but this black button makes this toy go," pointing to the two relevant buttons as he spoke. He pressed the incorrect black button. The button also did not activate the toy, and the puppet exclaimed "oh!" in surprise when nothing happened.

After each puppet pressed a button, the experimenter asked the test question: "[Child name], remember how I told you that one of my friends knows more about this toy? Can you tell me, which friend knows more?" Participants were asked to explain their answer. The inclusion questions were asked last, with children asked to match each puppet to the button he had pressed on the novel (inconsistent) toy: "[Child name], can you remind me, which one of my friends pressed this button [both puppets point to one button]? And which one of my friends pressed this button [both puppets point to the other button]?"

The identity of the puppet whose turn was first, and the button this agent pressed was always counterbalanced. Additionally, the side each puppet was presented on (left/right) was randomized.

6.3.2 Results and Discussion

Results were coded as in Experiment 1, with 26 participants excluded and replaced (see Participants). Overall, 61.1% of participants attributed knowledge to the puppet who pressed the black button, a proportion reliably higher than chance (44 of 72; 95% CI: 50 - 72.2). A logistic regression predicting performance based on age did not reveal a significant age difference ($\beta = 0.42$, p = .14). But while participants in all age groups preferred to attribute knowledge to the agent whose actions were inconsistent with his prior experience, only six-year-olds' preferences were robust. While 70.8% of six-year-olds judged that the agent who pressed the black button was more knowledgeable (17 of 24; 95% CI: 54.17 - 87.5), only 54% of four-year-olds (13 of 24; 95% CI: 33.33 - 75) and 58% of five-year-olds (14 of 24; 95% CI: 37.5 - 79.17) also made this judgment. In sum, although no age difference was obtained, only six-year-olds reliably judged that the agent whose failure was inconsistent with his prior experience had greater knowledge.

These findings suggest that children do not simply attribute knowledge to agents who show them something new. If they did, they should have performed at chance, as neither puppet taught children anything new. Instead, our results suggest that, by age six, children not only expect ignorant agents to act based on their prior knowledge, but also understand that knowledge runs along a continuum: agents can know more or less about any given topic. Thus, by age six, children attribute more knowledge to agents whose prior experience cannot explain their actions, even when these actions fail to fulfill their goal.

6.4 General Discussion

To successfully interact with others, we must understand what they know and believe, what they feel, and what they want. Children understand the link between mind and behavior early in life, inferring goals (Csibra, Gergely, Bíró, Koos, & Brockbank, 1999; Jara-Ettinger et al., 2016), beliefs (Onishi & Baillargeon, 2005; Rubio-Fernández & Geurts, 2013) and desires (Doan, Denison, Lucas, & Gopnik, 2015; Repacholi & Gopnik, 1997) from others' actions. Yet, while much work has shown that even young children have expectation about how knowledgeable agents should behave (Surian et al., 2007), less work has investigated whether children understand how ignorant agents might apply their prior knowledge to new situations.

Here we found that preschoolers expect ignorant agents to act based on their prior knowledge. When agents' past experience cannot explain their actions, children infer that these agents must have additional knowledge. In Experiment 1, five- and six-year-olds (but not four-year-olds) judged that an agent whose observable past experience could not explain his successful actions must've had additional knowledge. In Experiment 2, four- to six-year-olds (but only six-year-olds reliably) judged that an agent whose observable past experience could not explain his failure must've had some (incomplete) additional knowledge.

Our results show that, by age five, children expect past experiences to shape agents' beliefs and guide their actions in new situations. These results are consistent with related work, which suggests that children do not reliably link ignorance to specific outcomes (Friedman & Petrashek, 2009; German & Leslie, 2001; Ruffman, 1996).

These findings also suggest several broader implications. First, while we often talk about "knowing" or "not knowing," knowledge is not binary. People are rarely completely ignorant or completely knowledgeable. More frequently, knowledge lies along a continuum. In Experiment 2, six-year-olds succeeded in identifying which of two agents knew more, even when both agents were wrong. If children believe that agents can only be fully knowledgeable or fully ignorant, they may not have attributed even partial knowledge in this case (perhaps judging that any agent who is wrong is equally ignorant). The results of this experiment suggest that, by age six, children represent knowledge and ignorance as two poles of the epistemic continuum, leveraging their expectations about how prior experience should affect agents' actions to infer the extent of their knowledge.

Second, these findings provide insight into the development of children's epistemic inferences. While prior work has thoroughly investigated young children's ability to represent others' beliefs (e.g., Onishi & Baillargeon, 2005; Wellman et al., 2001), less research has investigated how children infer belief from action. In our tasks, children had to infer agents' beliefs from their actions. This required understanding that each agent pressed the button they believed would make the toy go, and considering what role their past experiences played in shaping these beliefs. Past work suggests that children infer knowledge from action via a naïve theory of knowledge: a set of expectations about how ignorant and/or knowledgeable agents should act (Aboody, Huey, & Jara-Ettinger, 2018). Our results are consistent with this account, demonstrating that across varied contexts, children can infer what others know or believe by observing their actions.

Our results also open avenues for future work. First, Experiment 2 shows that children do not simply attribute knowledge to agents who show them something new or surprising. However, other simple rules may explain participants' performance. For example, children may expect ignorant agents to act the same way they've acted in the past, without representing their knowledge or beliefs. In our studies, specifically, children may have solved the task by matching agents' current actions to their prior acts, licensing knowledge any time these acts were inconsistent. Future work can address this possibility by providing agents with knowledge, but not experience (e.g., by telling the puppets in Experiment 1 how the toy works but not allowing them to try it for themselves).

A second possibility is that children expect ignorant agents to try whatever is most reasonable, not in the context of agents' knowledge, but in the context of what children themselves think is reasonable. For example, children in our task could have assumed that the red button was the most obvious thing to try (regardless of agents'

past experiences), and attributed prior knowledge to any agent who rejected this obvious solution. While it is unclear whether children in fact find the red button to be the obvious solution in this task, future work can address this possibility by reversing Experiment 1, and introducing children to a training toy that works the same as the inconsistent toy. If children now attribute greater knowledge to the agent who presses the (more visually salient) red button, this would show that children do not just think that ignorance means trying the most perceptually obvious answer.

Third, in both experiments, puppets' actions differed, but so did their explanations of their actions. Namely, one agent said: "Hmm. Well, the red button made this [original] toy go, so the red button makes this toy go too," and the other said, "Hmm. Well, the red button made this [original] toy go, but this black button makes this toy go." Although only two words differed between explanations, it is possible that this could explain children's epistemic attributions in our task. Note, however, that this would be consistent with our account, showing that children attribute knowledge to those who explicitly reject past experience. In addition, if the linguistic cue guides children's inferences, this would be interesting in its own right—the difference between "so too" and "but" is subtle, and to our knowledge, little work has investigated how such words affect children's belief inferences. To identify whether these explanations were critical to children's inferences, future work will leave them out. If children make the same judgments, this would provide evidence that performance in this task did not hinge upon puppets' explanations.

Fourth, in Experiment 2, it is possible that children did not think both puppets were equally wrong. Conceptually, the puppet who pressed the black button may have been closer to being right (since he knew that one of the black buttons made the toy go). It is possible that children didn't consider whether agents' prior knowledge explained their actions, and instead simply attributed greater knowledge to the agent who was closer to being correct. While possible, this account does not explain children's success in Experiment 1. Furthermore, it is unclear how to operationalize what it means to be "closer" to being right in Experiment 2: while one agent was conceptually closer (pressing a black button), the other was physically closer (pressing the red button, which was right next to the correct black button). It is unclear how the magnitude of agents' errors may have guided children's inferences in the current task, but future work should investigate how this factor affects children's epistemic judgments.

Last, across both experiments, children's preferences strengthened with age (significantly in Experiment 1, and non-significantly in Experiment 2). Four-year-olds' failures in both experiments are consistent with prior work, which suggests that the ability to infer knowledge from behavior continues to develop between the ages of four and five (Aboody, Huey, & Jara-Ettinger, 2018). But while five-year-olds succeeded in Experiment 1, they were not reliably above chance in Experiment 2. Why might this be?

One possibility is that identifying a completely knowledgeable agent (Experiment

1) is easier than judging which agent has greater (but still incomplete) knowledge (Experiment 2). Furthermore, given that children may equate accuracy with knowledge (Brosseau-Liard & Birch, 2010; Ronfard & Corriveau, 2016), it might be harder for them to attribute knowledge in the face of a failure.

It is also possible that five-year-olds do attribute knowledge based on a rule (for example, attributing knowledge to agents who act in a surprising way). This could explain their weaker performance in Experiment 2, although it is unclear why four-year-olds would not have followed the same rule (which would have led to success in Experiment 1). It is possible that four-year-olds have no rule for inferring belief from knowledge, five-year-olds depend on a rule (e.g., knowledge = rejecting the obvious), and six-year-olds have a deeper understanding of how prior knowledge shapes beliefs. Finally, it is always possible that task demands affected children's performance, although this would fail to explain the difference in five-year-olds' performance across the two studies. Future work will address these possibilities to further clarify how children's epistemic intuitions emerge and develop.

In sum, across two experiments, we find evidence that young children have expectations for how prior knowledge is likely to shape people's beliefs and guide their behavior. We find that children use these expectations to infer what others know (or don't know) from their actions and that, by age five, children do not expect ignorant agents to act as blank slates; rather, they expect ignorant agents to leverage relevant prior knowledge when planning their actions. Altogether, our findings suggest that even young children may understand how ignorance begets belief and action.

6.4.1 Acknowledgments

We thank the Boston Children's Museum, the Peabody Natural History Museum, and the families who participated in this research. We thank Sarah Wong and Ivana Bozic for help with coding, and Lindsay Stoner for help with coding and data collection. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

Chapter 7

Discussion and Conclusions

Humans are uniquely social: we pass down our knowledge across generations, learn from those around us, and work together towards goals we may not even live to appreciate. Psychologists (not to mention philosophers) have long sought to uncover the foundations and development of our understanding of other minds. And yet, central puzzles remain unresolved. Specifically, researchers have long known that while young children readily reason about others' desires, goals, preferences and intentions, they struggle to reason about beliefs (see Wellman et al., 2001). This remains a classic puzzle in developmental psychology: why does epistemic reasoning lag behind other mental state reasoning capacities?

More recently, we have begun to learn more about the ways in which epistemic reasoning also lacks developmental cohesion. For example, the field has long assumed that an understanding of ignorance is more basic than an understanding of false belief: while a child may need to inhibit their own knowledge to reason about both mental states, only reasoning about a false belief also requires the child to realize that others' mental states do not have to reflect the true state of the world (e.g., an agent can have false mental representations, not just absent mental representations as in ignorance). However, recent empirical evidence challenges this view. For instance, even children who can represent false beliefs and use them to form explicit action predictions fail to understand how ignorance will lead an agent to act (Chen et al., 2015). And even five-year-olds, who generally pass false belief tasks, struggle to reason about ignorance in related tasks (Friedman & Petrashek, 2009; German & Leslie, 2001; Ruffman & Keenan, 1996).

To begin solving these puzzles, I proposed that our field may need to reconsider our approach to Theory of Mind, conceptualizing it not as a binary (have/have-not), but rather as a continuum of representational and computational capacities. I proposed that we could gain insight into the development of children's epistemic reasoning by better characterizing the representations and computations required—and that understanding how epistemic reasoning develops may allow us to better identify candidate causes for its relatively late emergence (e.g., conceptual change).

7.1 Chapters 2-6: A Review

In this thesis, I first proposed a novel theoretical account of epistemic inference. This account builds upon recent research characterizing human mental state reasoning (for a review, see Jara-Ettinger et al., 2016, 2020), but focuses specifically on characterizing inferences in the epistemic domain. Past work has begun to characterize how we might infer the precise content's of an agent's knowledge representation (specifying exactly what they know or believe; e.g., Baker et al., 2017; Jara-Ettinger et al., 2017). Our account advances this literature by successfully capturing adults' amorphous epistemic inferences (e.g., inferences over how *much* an agent knows or believes they can discover, without necessarily representing *what* they know precisely).

After validating our theoretical account (by formalizing it as a model, and showing that it well-captured adults' epistemic inferences across two experiments; Chapter 2), I investigated the developmental origins of adults' capacities. First, I tested whether preschoolers make epistemic inferences at all in situations where they cannot rely on simple heuristics, finding that by age five (and less reliably by age four) children do so (inferring that an agent who accurately says what's in a container before looking probably had prior knowledge, as compared to one who did the same after looking; Chapter 3).

Having established that preschoolers indeed make epistemic inferences even in situations where all external cues are matched, I leveraged my account of mature epistemic inference to begin breaking down what capacities might develop (and investigating how). This account has three main components: a) an expectation that others will trade-off their epistemic utilities, maximizing rewards and minimizing costs in the epistemic domain; b) an understanding of how knowledge affects an agent's probability of achieving their desired goal; and c) an understanding of how added knowledge might help.

In Chapter 4, I found that an expectation that others will trade-off epistemic utilities is early-emerging, already in place in our youngest four-year-old sample. This is consistent with recent research suggesting that even infants may show implicit evidence of such an expectation (Varga, Csibra, & Kovacs, 2021). In Chapter 5, however, I found that an understanding of the interaction between epistemic states and action outcomes was late-developing. It was not until age six that children evidenced an understanding of how different degrees of knowledge might affect an agent's probability of success. And in Chapter 6, I again found that not until age six did children reliably understand how added knowledge might affect an agent's chosen action plan (realizing that an agent who chose to reject what they had previously learned when activating a new toy, may have already known that this new toy worked differently).

7.2 What have we learned about development?

Taken together, these results begin to reveal both what expectations and computations may be early-emerging, and which may require time to emerge. Specifically, even young children expected others to trade-off their epistemic utilities (inferring that an agent who refused to incur a low cost to gain information must really not have needed it; and that an agent who incurred a high cost to seek information must have really wanted it). Thus, by age four (and perhaps earlier), children's mental state reasoning is already unified under an expectation that others will maximize their utilities.

However, an understanding of how different degrees of knowledge affect agents' actions (and their outcomes) was later-emerging. It was not until age six that children reliably understood how different epistemic states would likely lead agents to act. Younger children did not seem able to estimate how an agent's knowledge state would affect their ability to achieve their goals. These results are consistent with prior research suggesting that, although many domain-specific capacities mature during the preschool years (like executive function), children's models of other minds may also undergo conceptual development (Gopnik & Wellman, 1992, 1994; Wellman & Woolley, 1990; Wellman, 2014). This work helps pinpoint where development may be occurring: in being able to compute how different degrees of knowledge (e.g., full knowledge; partial knowledge; ignorance) will affect agents' actions (and thus their odds of success).

By leveraging a computational approach, we were able to systematically break down the components of a Theory of Mind (representations, action predictions, inferences). We identified a gap in the literature: the majority of past research in the epistemic domain has focused on the false belief task as a conservative test for Theory of Mind. This task requires children to represent and track what others know, and then use these representations to make action predictions. However, this task does not require children to make any epistemic inferences. Note also, that the false belief task requires children to make a relatively specific type of action prediction (usually, children have to leverage a false belief to decide where an agent will search; but they do not need to reason about how an agent's epistemic state might affect their odds of success in finding the item they desire).

In the work presented here, I find that these are precisely the capacities that are still developing throughout the preschool years. Thus, to understand the full richness of epistemic reasoning in our daily lives, I propose that we need rich and varied ways to study children's capacities. Looking back to classic research (which has relied on a variety of epistemic reasoning tasks, not just the false belief task; e.g., Wellman & Liu, 2004; Wellman, 1992), I argue that we must again diversify the tasks we use to study epistemic reasoning—now including measures of epistemic inference as well as epistemic action prediction. And looking to current research investigating mental state reasoning through a formal lens, I argue that we ought to ground our investigations in formal frameworks, in order to systematically break

down the components of the capacities we seek to study, and produce precise, testable predictions from theoretical-level accounts.

7.3 Open questions

7.3.1 Proposing a new and updated Theory of Mind battery

This work also opens many questions for future research. First, I found that four- and five-year-olds often performed at chance when asked to reason about how knowledge might affect an agent's actions (and probability of success). While chance performance could reflect a general inability, the same overall pattern could result if some younger children systematically performed above chance, and some performed systematically below. Indeed, preschoolers' performance on the false belief task begins at below-chance levels (Wellman, 1992); could it be the case that some younger children are capable of making epistemic inferences, but their performance is masked by children who cannot? Here, I have presented a collection of disparate tasks, usually including only a single forced-choice test question (intended to reduce task demands, as epistemic reasoning tasks can be challenging for young children). Although many of these tasks conceptually replicate those that came before (and generally result in a similar developmental trajectory), these tasks cannot reveal why some younger children succeeded and some failed. However, an updated Theory of Mind battery (a la Wellman & Liu, 2004), where the same children complete a variety of different mental state reasoning tasks (including epistemic inference tasks) could answer this question. Furthermore, such a battery would have the potential to reveal the impact of task demands on the developmental trajectory I identified. I hope to investigate these questions in future research.

7.3.2 Can children represent graded or amorphous epistemic states?

Second, in Chapter 1, I began by pointing out that the majority of past research in the epistemic domain has focused on questions of representation and action prediction. By studying epistemic inferences as well, I sought to go beyond past work, to better understand the full developmental trajectory of epistemic reasoning. While I hope I have done so, my results have also led me back to where our field began, considering basic questions of representation. The majority of past research in the epistemic domain, including traditional false belief tasks, has focused on cases where an agent either does or does not know about one specific thing (e.g., the location of an object). This leaves open the question of how children represent what others know in more complex situations: for instance, when representing partial knowledge. In some cases, these richer epistemic representations could simply take the form of representing which specific things agents do or do not know (e.g., Anne knows there's a

sandwich in her lunchbox, and does not know there is also chocolate). But as adults, we are able to represent graded epistemic states, and reason over them flexibly (e.g., Chapter 2). When communicating, we may need to know what specific pieces of information are in the common-ground; but we may also need to represent approximately how much our interlocutor knows to teach at the right level, or describe a concept in a comprehensible way. While young children grasp basic concepts like expertise, which communicate that an agent is generally knowledgeable about a domain without enumerating exactly what they know (Lutz & Keil, 2002), it is not entirely clear how they do so (and whether they can do the same in contexts where domain knowledge does not apply; such as the inferences adults made in Chapter 2). Thus, in future work, I hope to investigate how children represent more graded knowledge states, which also requires better understanding how children represent knowledge and ignorance in more varied contexts.

7.3.3 Capturing the breadth of epistemic reasoning

Finally, in this thesis, I have focused primarily on characterizing the development of epistemic inference. However, this comprises only one slice of epistemic reasoning. To fully understand epistemic reasoning, we need to understand not only how capacities like epistemic inference develop (Chapters 2-6); we also need to understand how epistemic reasoning functions in everyday life. For instance, to learn effectively, we need to understand what our teachers know (about the world, and about ourselves; Asaba & Gweon, 2020; Bass, Mahaffey, & Bonawitz, 2021; Bass et al., 2022). To explore, we may need a sense of what *we* know, and what is knowable. To teach effectively, we need a good sense for what our learner knows or believes (e.g., Aboody, Velez-Ginorio, Santos, & Jara-Ettinger, 2018). To communicate, we must monitor what's in common ground and what is not. To cooperate, we need to be confident everybody involved knows their role. To morally evaluate, we need to consider whether someone knowingly transgressed or made an innocent mistake. To navigate the world, we must have some sense of how knowledge passes between individuals (Aboody, Yousif, Sheskin, & Keil, 2022; Yousif, Aboody, & Keil, 2019), or within and across social groups. Do each of these capacities rely on different expectations or computations? Or do we apply the same computations across epistemic reasoning tasks, even ones that vary widely in their domain? In future work, I would like to explore these questions, better understanding the system (or systems) that enable us to engage in epistemic reasoning, broadly construed.

7.4 Conclusions

Taken together, the work presented here has advanced our understanding of epistemic reasoning in two main respects. First, I provided a novel theoretical account of epistemic inference, formalized as a computational model, that makes precise and testable

predictions. Second, after validating this account with adults, I used its predictions to investigate the development of epistemic inference across four sets of experiments. I found that four-year-olds and adults rely on the same basic assumptions to infer what others know: expecting others to trade off information's cost and reward. My work also suggests that children's understanding of other minds does undergo conceptual development in the preschool years, and helps pinpoint what precisely is developing: an ability to estimate how knowledge will affect an agent's actions (and their likely outcomes). This work lays a strong theoretical and empirical foundation for future research to continue investigating the fundamental building blocks and developmental trajectory of epistemic inference—as well as epistemic reasoning more broadly.

References

Aboody, R., Denison, S., & Jara-Ettinger, J. (2021). Children consider the probability of random success when evaluating knowledge. In *Cogsci.*

Aboody, R., Huey, H., & Jara-Ettinger, J. (2018). Success does not imply knowledge: Preschoolers believe that accurate predictions reveal prior knowledge, but accurate observations do not. In *Cogsci.*

Aboody, R., Velez-Ginorio, J., Santos, L., & Jara-Ettinger, J. (2018). When teaching breaks down: Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces. In *Cogsci.*

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. (2022). Says who? children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*.

Aboody, R., Zhou, C., Flowers, M., & Jara-Ettinger, J. (2019). Ignorance= doing what is reasonable: Children expect ignorant agents to act based on prior knowledge. In *Cogsci* (pp. 1297–1303).

Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Development*, *92*(5), 1919–1931.

Asaba, M., & Gweon, H. (2020). Learning about others to learn about the self: Early reasoning about the informativeness of others' praise. In *Psychological perspectives on praise* (pp. 67–74). Routledge.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10.

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

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* Oxford University Press.

Bass, I., Bonawitz, E., Hawthorne-Madell, D., Vong, W. K., Goodman, N. D., & Gweon, H. (2022). The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition*, *222*, 104999.

Bass, I., Mahaffey, E., & Bonawitz, E. (2021). Do you know what i know? children use informants' beliefs about their abilities to calibrate choices during pedagogy. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Begus, K., Gliga, T., & Southgate, V. (2016). Infants' preferences for native speakers

are associated with an expectation of information. *Proceedings of the National Academy of Sciences*, *113*(44), 12397–12402.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 153–164.

Bernard, S., Castelain, T., Mercier, H., Kaufmann, L., Van der Henst, J.-B., & Clément, F. (2016). The boss is always right: Preschoolers endorse the testimony of a dominant over that of a subordinate. *Journal of Experimental Child Psychology*, *152*, 307–317.

Birch, S. A., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369.

Bohn, M., & Köymen, B. (2018). Common ground and development. *Child Development Perspectives*, *12*(2), 104–108.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bonawitz, E., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Bradmetz, J., & Schneider, R. (1999). Is little red riding hood afraid of her grandmother? cognitive vs. emotional response to a false belief. *British Journal of Developmental Psychology*, *17*(4), 501–514.

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology*, *52*(1), 9.

Bridgers, S., Gweon, H., Bretzke, M., & Ruggeri, A. (2018). How you learned matters: The process by which others learn informs young children's decisions about whom to ask for help. In *Cogsci.*

Brosseau-Liard, P. E., & Birch, S. A. (2010). 'i bet you know more and are nicer too!': What children infer from others' accuracy. *Developmental Science*, *13*(5), 772–778.

Brosseau-Liard, P. E., & Poulin-Dubois, D. (2014). Sensitivity to confidence cues increases during the second year of life. *Infancy*, *19*(5), 461–475.

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? contributions of inhibitory control and working memory. *Infant and Child Development: An International Journal of Research and Practice*, *11*(2), 73–92.

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development*, *69*(3), 672–691.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*(6), 657–674.

Chen, Y., Su, Y., & Wang, Y. (2015). Young children use the "ignorance= getting it wrong" rule when predicting behavior. *Cognitive Development*, *35*, 79–91.

Chow, V., Poulin-Dubois, D., & Lewis, J. (2008). To see or not to see: Infants prefer to follow the gaze of a reliable looker. *Developmental Science*, *11*(5), 761–770.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Corriveau, K., & Harris, P. L. (2009a). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426–437.

Corriveau, K., & Harris, P. L. (2009b). Preschoolers continue to trust a more accurate informant 1 week after exposure to accuracy information. *Developmental Science*, *12*(1), 188–193.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 447–458.

Csibra, G., Gergely, G., Bíró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason'in infancy. *Cognition*, *72*(3), 237–267.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5-and 6-month-olds. *Developmental Psychology*, *49*(2), 243.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Diesendruck, G., Salzer, S., Kushnir, T., & Xu, F. (2015). When choices are not personal: The effect of statistical and social cues on children's inferences about the scope of preferences. *Journal of Cognition and Development*, *16*(2), 370–380.

Doan, T., Denison, S., Lucas, C. G., & Gopnik, A. (2015). Learning to reason about desires: An infant training study. In *Cogsci*.

Doan, T., Friedman, O., & Denison, S. (2018). Beyond belief: The probability-based notion of surprise in children. *Emotion*, *18*(8), 1163.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences*, *12*(7), 248–253.

Einav, S., & Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable informants from merely accurate informants. *Psychological Science*, *22*(10), 1250–1253.

Friedman, O., & Petrashek, A. R. (2009). Children do not follow the rule "ignorance means getting it wrong". *Journal of Experimental Child Psychology*, *102*(1), 114–121.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*(4), 483–527.

Fusaro, M., Corriveau, K. H., & Harris, P. L. (2011). The good, the strong, and the accurate: Preschoolers' evaluations of informant attributes. *Journal of Experimental Child Psychology*, *110*(4), 561–574.

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

German, T. P., & Leslie, A. M. (2001). Children's inferences from 'knowing'to 'pretending'and 'believing'. *British Journal of Developmental Psychology*, *19*(1), 59–83.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind and Language*, *7*, 145–171.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (p. 257-293).

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524–1524.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Hala, S., & Russell, J. (2001). Executive control within strategic deception: A window on early cognitive development? *Journal of Experimental Child Psychology*, *80*(2), 112–141.

Harris, P. L. (2012). *Trusting what you're told*. Harvard University Press.

Horn, S. S., Ruggeri, A., & Pachur, T. (2016). The development of adaptive decision making: Recognition-based inference in children and adolescents. *Developmental Psychology*, *52*(9), 1470.

Hughes, C. (1998). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology*, *34*(6), 1326.

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

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

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

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

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

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jern, A., & Kemp, C. (2014). Reasoning about social choices and social relationships. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

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

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

Josephs, M., Kushnir, T., Gräfenhain, M., & Rakoczy, H. (2016). Children protest moral and conventional violations more when they believe actions are freely chosen. *Journal of Experimental Child Psychology*, *141*, 247–255.

Kachel, U., Svetlova, M., & Tomasello, M. (2018). Three-year-olds' reactions to a partner's failure to perform her role in a joint commitment. *Child Development*, *89*(5), 1691–1703.

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of southgate, senju and csibra (2007). *Royal Society open science*, *8*(8), 210190.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Kinzler, K. D., Corriveau, K. H., & Harris, P. L. (2011). Children's selective trust in native-accented speakers. *Developmental Science*, *14*(1), 106–111.

Koenig, M. A. (2012). Beyond semantic accuracy: Preschoolers evaluate a speaker's reasons. *Child Development*, *83*(3), 1051–1063.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694–698.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277.

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology*, *52*(1), 31.

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

Landrum, A. R., & Mills, C. M. (2015). Developing expectations regarding the boundaries of expertise. *Cognition*.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294.

Liu, D., Gelman, S. A., & Wellman, H. M. (2007). Components of young children's trait understanding: Behavior-to-trait inferences and trait-to-behavior predictions. *Child Development*, *78*(5), 1543–1558.

Liu, S., Cushman, F., Gershman, S., Kool, W., & Spelke, E. S. (2019). Hard choices: Children's understanding of the cost of action selection. In *Cogsci* (pp. 671–6677).

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42.

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

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

Luchkina, E., Morgan, J. L., Williams, D. J., & Sobel, D. M. (2020). Questions can answer questions about mechanisms of preschoolers' selective word learning. *Child Development*, *91*(5), e1119–e1133.

Luchkina, E., Sobel, D. M., & Morgan, J. L. (2018). Eighteen-month-olds selectively generalize words from accurate speakers to novel contexts. *Developmental Science*, *21*(6), e12663.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084.

MacLaren, R., & Olson, D. (1993). Trick or treat: Children's understanding of surprise. *Cognitive Development*, *8*(1), 27–46.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* MIT Press.

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838.

Nurmsoo, E., & Robinson, E. J. (2009a). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. *Child Development*, *80*(1), 23–27.

Nurmsoo, E., & Robinson, E. J. (2009b). Identifying unreliable informants: Do children excuse past inaccuracy? *Developmental Science*, *12*(1), 41–47.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*(2), 659–677.

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

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216.

Pesowski, M. L., Denison, S., & Friedman, O. (2016). Young children infer preferences from a single action, but not if it is constrained. *Cognition*, *155*, 168–175.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ... Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, *44*.

Poulin-Dubois, D., Brooker, I., & Polonia, A. (2011). Infants prefer to imitate a reliable person. *Infant Behavior and Development*, *34*(2), 303–309.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50.

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

Richardson, E., & Keil, F. (2020). Children use agents' response time to distinguish between memory and novel inference. In *Cogsci.*

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1–12.

Robinson, E. J., Butterfill, S. A., & Nurmsoo, E. (2011). Gaining knowledge via other minds: Children's flexible trust in others as sources of information. *British Journal of Developmental Psychology*, *29*(4), 961–980.

Robinson, E. J., & Whitcombe, E. (2003). Children's suggestibility in relation to their understanding about sources of knowledge. *Child Development*, *74*(1), 48–62.

Ronfard, S., & Corriveau, K. H. (2016). Teaching and preschoolers' ability to infer knowledge from mistakes. *Journal of Experimental Child Psychology*, *150*, 87–98.

Rubio-Fernández, P. (2019). Memory and inferential processes in false-belief tasks: An investigation of the unexpected-contents paradigm. *Journal of Experimental Child Psychology*, *177*, 297–312.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27–33.

Ruble, D. N., & Dweck, C. S. (1995). Self-conceptions, person conceptions, and their development. *Review of Personality and Social Psychology*, *15*, 109–139.

Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? evidence from their understanding of inference. *Mind & Language*, *11*(4), 388–414.

Ruffman, T., & Keenan, T. R. (1996). The belief-based emotion of suprise: The case for a lag in understanding relative to false belief. *Developmental Psychology*, *32*(1), 40.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159.

Russell, J. (1996). Agency: Its role in mental development hove. *England: Erlbaum*.

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

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*, *30*(11), 1561–1572.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045.

Scott, R. M. (2017). Surprise! 20-month-old infants understand the emotional consequences of false beliefs. *Cognition*, *159*, 33–47.

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? causal inference and the development of social attribution. *Child Development*, *84*(2), 443–454.

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779.

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

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586.

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, *115*(34), 8491–8498.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*.

Varga, B., Csibra, G., & Kovacs, A. (2021). Infants' interpretation of information-seeking actions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Velez-Ginorio, J., Siegel, M. H., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires. In *Cogsci.*

Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin & Review*, *24*(5), 1538–1547.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

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

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

Wellman, H. M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, *5*(1), 33–38.

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

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*(3), 245–275.

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

Wu, Y., & Schulz, L. E. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*, *89*(2), 649–662.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.

Yousif, S. R., Aboody, R., & Keil, F. C. (2019). The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science*, *30*(8), 1195–1204.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology*, *1*(2), 198–226.

Zmyj, N., Buttelmann, D., Carpenter, M., & Daum, M. M. (2010). The reliability of a model influences 14-month-olds' imitation. *Journal of Experimental Child Psychology*, *106*(4), 208–220.