Communication in Action: Planning and Interpreting Communicative Demonstrations

Mark K. Ho
Princeton University

Fiery Cushman
Harvard University

Michael L. Littman
Brown University

Joseph L. Austerweil
University of Wisconsin-Madison

Theory of mind enables an observer to interpret others' behavior in terms of unobservable beliefs, desires, intentions, feelings, and expectations about the world. This also empowers the person whose behavior is being observed: By intelligently modifying her actions, she can influence the mental representations that an observer ascribes to her, and by extension, what the observer comes to believe about the world. That is, she can engage in intentionally *communicative demonstrations*. Here, we develop a computational account of generating and interpreting communicative demonstrations by explicitly distinguishing between two interacting types of planning. Typically, *instrumental* planning aims to control states of the environment, whereas *belief-directed* planning aims to influence an observer's mental representations. Our framework extends existing formal models of pragmatics and pedagogy to the setting of value-guided decision-making, captures how people modify their intentional behavior to show what they know about the reward or causal structure of an environment, and helps explain data on infant and child imitation in terms of literal versus pragmatic interpretation of adult demonstrators' actions. Additionally, our analysis of belief-directed intentionality and mentalizing sheds light on the socio-cognitive mechanisms that underlie distinctly human forms of communication, culture, and sociality.

Keywords: communication, problem solving, pragmatics, planning, social learning

Pre-print December 2020—to appear in JEP:General.

Introduction

Communicating often requires demonstration. Imagine teaching a child to tie her shoes with words alone, or by simply showing her the finished product—this is unlikely to work. Instead, we must show her.

Because communicative demonstrations are essential for humans, they are also routine. They occur when we coordinate (Clark, 2005), cooperate (Jordan, Hoffman, Bloom, & Rand, 2016), create novel signs (Scott-Phillips, Kirby, & Ritchie, 2009), and control low-level motor behaviors during interaction (Wolpert, Doya, & Kawato, 2003; Pezzulo et al., 2019). Developmental psychologists, especially, emphasize the importance of communicative demonstrations. This is because such social interactions enable infants and children to learn a range of useful behaviors and representations, including action types, subgoals, tool functions, causal structure, and normative concepts (Brand, Baldwin, & Ashburn, 2002; Brugger, Lariviere, Mumme, & Bushnell, 2007; Southgate, Chevallier, & Csibra, 2009; Király, Csibra, & Gergely, 2013; Hernik & Csibra, 2015; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler, Schmidt, Bürgel, & Tomasello, 2015; Sage & Baldwin, 2011; Hoehl, Zettersten, Schleihauf, Grätz, & Pauen, 2014).

How do communicative demonstrations work? What cognitive processes support generating and interpreting demonstrations, as well as related communicative actions such as gestures (Cartmill, Beilock, & Goldin-Meadow, 2012) and depictions (Clark, 2016)? Intuitively, demonstrative shoe-tying is very similar to ordinary shoe-tying, but also importantly distinct. When demonstrating we tie our shoes slowly, with exaggerated motions, pausing at and repeating certain key actions. When watching a communicative demonstration these distinctive features serve as important clues, revealing which parts of a sequence of actions are essential and which are merely incidental.

Communicative demonstration takes an ordinary act with its ordinary purpose and builds something richer on top of it. It depends upon a shared understanding between the actor and the observer: That the actor intends not just perform the ordinary action, but also to convey something about it. This shared understanding allows each pause, repetition and exaggeration to carry special significance.

Our goal is to explain how this works and why it is so important. Following others who have studied how we communicate with our actions (Sperber & Wilson, 1986; Tomasello, Carpenter, Call, Behne, & Moll, 2005; Csibra & Gergely, 2009; Tomasello, 2010; Clark, 2016), we draw inspiration from a different medium of human communication: language. Just as demonstration layers communication on top of simple goal-directed action, language layers pragmatic inference on top of simple literal meaning. This analogy is central to our approach.

More specifically, contemporary accounts of language emphasize that speakers' words are not only chosen according to their conventional semantic meaning, but also according to a model of how they will be interpreted by the listener. Listeners, in turn, often reason about utterances in light of these goals. For example, consider *scalar implicature* (Spector, 2007; Frank & Goodman, 2012): When a friend says to you, "I ate some of the pizza in the refrigerator," how does this change your beliefs about the leftover pizza? Although the literal meaning of the statement is consistent with them having eaten *all* of the pizza, in an everyday context the statement implies that *not all* of the pizza was eaten; there is some left over. This is because you both know that your partner has an intention to inform you and that if they had wanted to inform you that was no pizza left, they would have said they had eaten *all* of the pizza. These pragmatic aspects of language use and comprehension have been extensively studied (Grice, 1957; Horn, 1984; Sperber & Wilson, 1986; Clark, 1996; Levinson, 2000).

We aim to show that communicative demonstration operates by similar logic. When demonstrating an action, we do not just orient our behavior around its ordinary goal (analogous to a "literal" semantics), but also around an understanding of the actor's communicative intent (analogous to "figurative", or pragmatic meaning). This general theme has been been examined by researchers in a number of disciplines, including linguistics (Clark, 2005, 2016), comparative psychology (Tomasello, 2010), and developmental psychology (Csibra & Gergely, 2009). Here, we precisely characterize the cognitive mechanisms underlying communicative demonstrations within a general mathematical framework of probabilistic inference and decision-making. In particular, we build on the general ideas developed for cooperative communication (Shafto, Goodman, & Frank, 2012; Shafto, Goodman, & Griffiths, 2014), game-theoretic experimental pragmatics (Franke, 2009), and rational speech-act theory (Goodman & Frank, 2016) and adapt them to the domain of actions. To accomplish this, we integrate them with a distinct set of ideas developed to model communication in the context of goal-directed planning and decision-making (Newell & Simon, 1972; Sutton & Barto, 1998; Dayan & Niv, 2008). This marriage of formal tools for modeling decision-making and pragmatic inference is at the heart of our proposal.

By drawing on ideas from pedagogy and pragmatics, our computational approach helps answer two key questions about how communicative demonstrations work. First, what allows non-linguistic actions to have meaning? The literal semantics of words are essential to creating their additional pragmatic meaning, but actions (e.g., tying one's shoes) must derive their meaning differently. Second, how can actions literally *do* things as well as figuratively *show* things? A demonstrator must be able to anticipate and reason about the literal effects of their actions (e.g., how they attain a secure bow) as well as their communicative effects (e.g., how they convey how to attain a secure bow). Meanwhile, an observer must be able to determine whether actions are merely literal or also communicative, and if so, what they are communicating. Our aim is to answer these questions within a single computational framework and test their empirical predictions.

Two kinds of action and action interpretation

To characterize communicative demonstrations, we draw on a distinction between two types of action (Shafto, Goodman, & Frank, 2012): instrumental and belief-directed. Instrumental actions are prototypically aimed at solving physical problems or accomplishing physical goals in an environment. For example, when someone is riding a bicycle, they pedal with their legs, causing the wheels to turn. These actions are taken instrumentally to achieve a desired outcome, such as quickly reaching a destination.

An agent's actions can be interpreted as instrumental. Indeed, this is the ordinary manner of interpreting actions. It includes reasoning about a range of mental states about the environment including others' goals (Gergely, Nádasdy,

Contact: mho@princeton.edu. The authors would like to thank Patrick Shafto, Michael J. Frank, Otávio Mattos, Rachit Dubey and the anonymous reviewers for comments on earlier versions of this work.

Web link for materials: https://github.com/markkho/comdem-data-code

Preliminary versions of this work have been presented at Neural Information Processing Systems Conference (Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016) as well as the Annual Meeting of the Cognitive Science Society (Ho, Littman, Cushman, & Austerweil, 2018).

This research was performed using the compute resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

JLA was supported by the Office of the CVGRE at UW-Madison with funding from the WARF. FAC was supported by grant N00014-19-1-2025 from the Office of Naval Research and grant 61061 from the John Templeton Foundation.

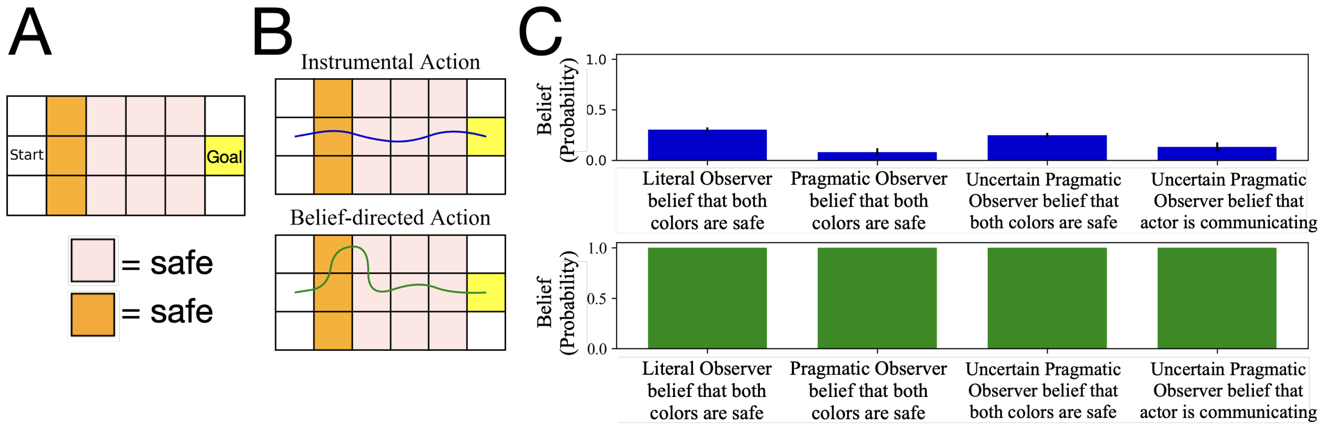

Figure 1. Example predictions. Our account makes predictions about how instrumental actions (e.g., riding a bicycle) relate to communicative demonstrations (e.g., showing someone you can ride a bicycle without holding onto the handlebars). (A) An example domain in which one must navigate from the start location to the goal (+10 points) while avoiding tiles that are dangerous. Orange and pink tiles could be dangerous (-1 point) or safe (0 points), but in this particular case both colors are safe. (B) Examples of instrumental versus belief-directed action predicted by our models. Top: The optimal instrumental sequence of actions is to head straight towards the goal. Note that this is the case regardless of which tiles are safe or dangerous. Bottom: Optimal belief-directed action sequences (i.e., communicative demonstrations) involve visiting extra orange and pink tiles in order to show that they are safe. (C) Inverse planning inferences drawn by three types of model observers when presented with the example demonstrations. A *literal observer* reasons about actions in terms of instrumental planning. A *pragmatic observer* reasons about actions in terms of both instrumental planning and an intention to convey information about the world. An *uncertain pragmatic observer* reasons jointly about the target (e.g., that orange and pink tiles are safe in this case) and whether the actor has communicative intent. Top: Mean final beliefs of simulated observer models given instrumental action sequences (blue; n = 50; error bars are standard errors). All three observer models draw relatively weak inferences about the target, and the uncertain pragmatic observer correctly infers that demonstrators do not have communicative (belief-directed) goals. Bottom: Mean final inferences for observer models given belief-directed action sequences (green; n = 50). All three observers infer the target with high probability and the uncertain pragmatic observer correctly infers that the demonstrations are communicative.

Csibra, & Bíró, 1995; Baker, Saxe, & Tenenbaum, 2009), action costs (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015), and false beliefs about the world (Southgate, Senju, & Csibra, 2007). In these typical applications of "theory of mind"(Premack & Woodruff, 1978) an actor's behavior is interpreted as instrumental.

In principle, human social learning could depend exclusively on actors performing instrumental actions and observers interpreting them as such. And, these are certainly necessary features of any account of human social learning. In practice, however, they cannot be sufficient.

First, true communicative demonstration involves more than mere instrumental action. Like linguistic utterances (Grice, 1957) but unlike typical, instrumental actions, communicative demonstrations are taken in order to affect another's cognitive state—for instance, to teach them. A father demonstrating how to tie shoes does not organize his action around the sole goal of shoe-tying. Rather, he has the additional goal of conveying various aspects of this general skill. His actions are directed not just towards an object, but also towards her thoughts and beliefs. In short, his actions are both instrumental and belief-directed.

Second, demonstrations are often interpreted differently that ordinary actions. For instance, research on observational learning has shown that infants and children are highly sensitive to "ostensive cues", which are cues generated by an adult that make it clear that they want their actions to be observed. The presence or absence of ostensive cues can have a radical effect on how learners interpret the very same sequence of actions (Király et al., 2013; Butler et al., 2015; Hernik & Csibra, 2015). How does interpreting an actor that clearly "wants to be seen" differ from one that does not care whether they are seen? Intuitively, wanting to be seen does not inherently change the space of possible instrumental intentions. Rather, it announces the presence of possible belief-directed intentions (e.g., "I want you to see this so that you can infer something"), which a learner may then reason about. In other words, whereas intentional actions are interpreted as "literally" instrumental, ostensive cues nudge an observer towards also *pragmatically* interpreting actions as resulting from belief-directed goals.

Our goal is to capture the mechanics of these different types of action and interpretation. We start by formalizing the distinction between instrumental action and belief-directed action and their interaction in communicative demonstrations. We then report two sets of new experiments designed to test the key features of this formal account. Next, we show how the formal account also captures key elements of prior studies on infant social learning. Finally, we show how our account extends existing computational approaches to language pragmatics and pedagogy, how it connects with a number of active topics in social cognition, and what it suggests for future research on the cognitive processes underlying human communication.

General Methods

Our goal is to understand how people produce, and others learn from, communicative demonstrations. In particular, our aim is to provide a normative, computational-level account of these processes in the sense of predicting how communicative demonstrations should be performed and interpreted by rational agents (Marr, 1982; Anderson, 1990). In this section, we describe the computational framework that spells out these assumptions and structures our investigation.

It is organized around two key ideas, which together comprise this work's main contributions. The first idea is that demonstrators and observers are engaged in *pragmatic reasoning* that is grounded in interpreting actions as instrumental. In other words, each explicitly models the problem of sending and receiving maximally informative signals. To formalize this, we borrow from prior models of pedagogical and pragmatic reasoning (Shafto et al., 2014; Frank & Goodman, 2012; Rafferty, Brunskill, Griffiths, & Shafto, 2016; Sperber & Wilson, 1986; Grice, 1957).

Specifically, we analyze pragmatic meaning as emerging from recursive social reasoning (Camerer, Ho, & Chong, 2004). In this approach, a model begins by specifying a "literal" instrumental actor who chooses an action without modeling the mental state reasoning of an observer; the observer then models this choice by reasoning about the actor's mental states; the actor then chooses an action that maximizes the probability of the observer drawing the correct inference; the observer then models the actor as such; and so on. In theory such "cognitive hierarchies" could proceed *ad infinitum*. In practice, they attain their predictive power within a few layers of recursive mentalizing (Camerer et al., 2004).

The second idea is that communicative demonstrations involve reasoning about actions as both *instrumental* and *belief-directed*. For example, a father showing his daughter how to tie shoes both wants her shoes to be tied (an instrumental goal) and also wants his daughter to learn how to tie shoes (a belief-directed goal). Not only does this require balancing two types of (potentially competing) goals, but it requires reasoning about two distinct types of causal effects: how actions influence the environment as well as how they also influence an observer's mental state. Meanwhile, an observer must be able to interpret actions in terms of these different levels. To characterize these planning and inference processes, our approach marries insights from the study of goal-directed planning (Dayan & Niv, 2008; Newell & Simon, 1972; Puterman, 1994) and theory of mind (Dennett, 1987; Malle, 2008; Gergely & Csibra, 2003; Baker et al., 2009).

These two key ideas—grounding the pragmatics of communicative action in instrumental action, and planning actions over a model of instrumental and belief-directed effects—can be combined in a straightforward and productive manner. Specifically, we begin by defining a form of instrumental action production and observation. We then allow belief-directed goals to structure pragmatic reasoning that arises as the next level up of cognitive hierarchy. In this manner a relatively simple and traditional planning problem of attaining an instrumental goal (e.g., catching a fish) "grounds" the pragmatic inferences that structure the additional and more complex planning problem of attaining a belief-directed goal (e.g., teaching a person to fish).

Instrumental planning and literal action interpretation

At the first level of a cognitive hierarchy we model an actor attempting to accomplish an instrumental goal—i.e., to perform a task without regard for an observer. In order to do this she must engage in planning, which involves reasoning about what actions and associated consequences best achieve her goals (Figure 2A). We suppose an actor has a model of how her actions will affect the environment, W, and instrumental goals expressed in terms of utilities, G_I. We denote an instrumental plan as π_I. This can be thought of as computing the steps of a procedure (e.g., "First bring the flour down from the shelf. Then get a spoon. Take out 2 cups." etc.). When an instrumental demonstrator acts out a plan, this produces a sequence of events, which includes her physical actions and their consequences (e.g., reaching for the flour, moving it from the shelf to the counter, picking up a spoon, etc.). Given a model and goals, an intentional agent plans and then acts. We can express the probability of a demonstration D (i.e., a particular sequence of actions and consequences):

$$P(D \mid W, G_I) = \sum_{\pi_I} p(D \mid \pi_I, W) P(\pi_I \mid W, G_I) \qquad (1)$$

In Equation 1, $P(\pi_I \mid W, G_I)$ expresses the output of a rational planning process, while $p(D \mid \pi_I, W)$ expresses how an actor's planned actions interact with the environment. This kind of model-based action selection is widely explored in the literature on value-guided decision-making (Dayan & Niv, 2008). An agent who plans and acts in this manner is oblivious to the fact that anyone may be observing her. We describe this model not because it describes an agent engaged in communicative demonstration—it does not!—but rather because it grounds successive levels of hierarchical mental state inference that we are interested in. The next step is to model an observer who attempts to learn from an actor's behavior by assuming that the actor is purely instrumental (i.e., by assuming the model summarized in Equation 1).

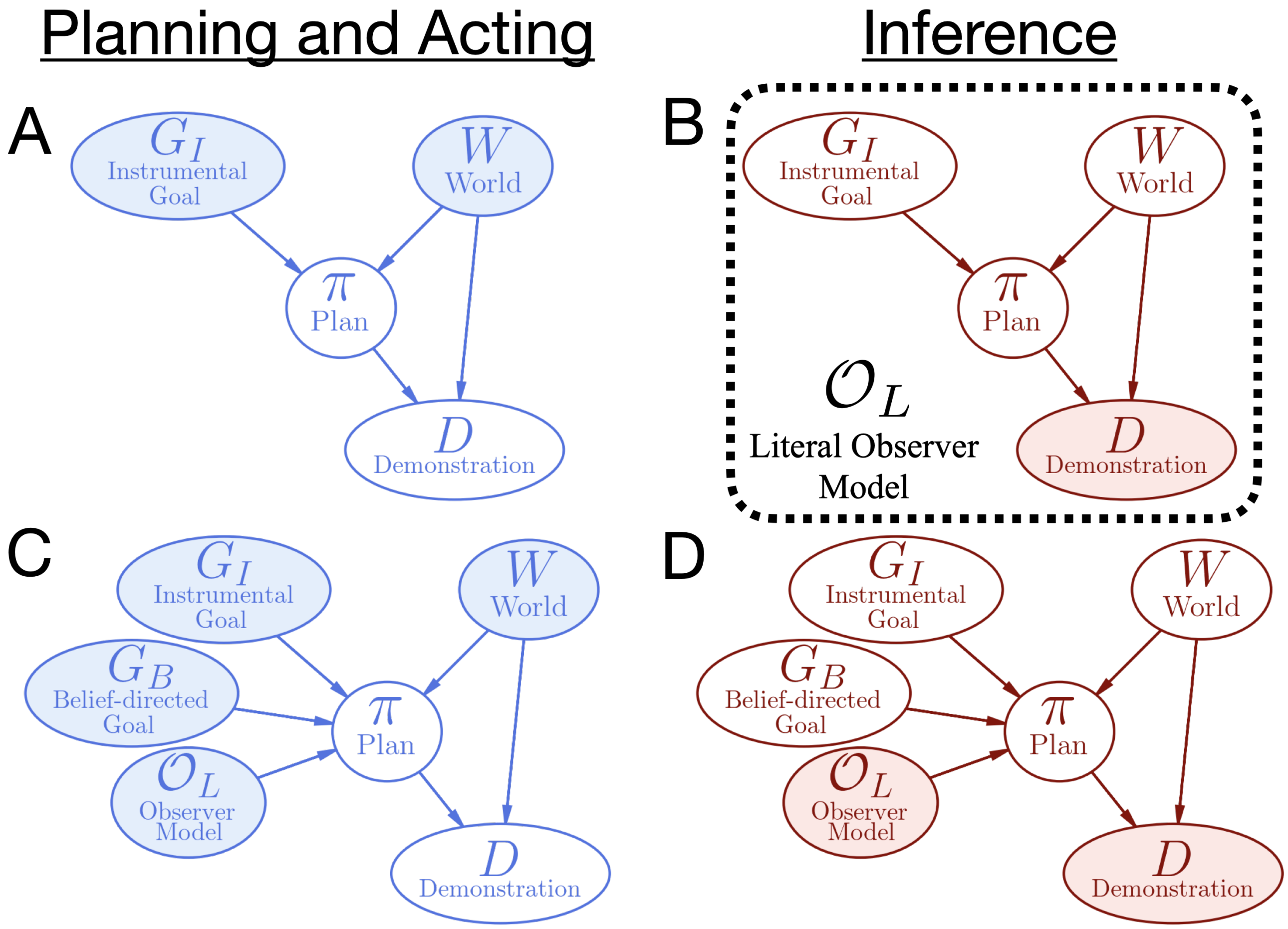

Figure 2. Bayesian Network Diagrams (Pearl, 1988) of the relationships between instrumental and belief-directed planning and inference. Nodes represent variables in the model, edges indicate causal influence, and highlighted nodes are directly observed or assumed by either the demonstrator (blue) or observer (red). (A) During *instrumental action*, an actor generates a plan, π, given instrumental goals, G_I, and the world, W. Enacting the plan in the world results in a demonstration, D. (B) A *literal observer* inverts the model of instrumental planning (Baker et al., 2009). This allows them to observe a demonstration D and draw inferences about instrumental goals G_I and the world W by reasoning about instrumental planning π. (C) During a *communicative demonstration*, an actor's plan π is further determined by belief-directed goals G_B and their model of a literal observer's inference process O_L. Note that O_L is derived directly from the inferences represented in panel B (see main text for details). (D) Given an observed demonstration D and knowledge of how a literal observer would interpret it (i.e., O_L), a *pragmatic observer* reasons about the world W via inferences about belief-directed planning π. Together, panels C and D illustrate a key aspect of the model: Shared understanding of instrumental action and its interpretation (i.e., the literal observer model) provides a way for a communicative demonstrator and pragmatic observer to coordinate the meaning of demonstrations.

Specifically, to an observer, a demonstrator's knowledge of the world W and her instrumental goals G_I are hidden. However, by reasoning about the relationship between a demonstrator's plans and actions, he can infer her beliefs, desires and intentions (Figure 2B). For example, when you observe your neighbor pull her bicycle out of their garage and then ride it towards work for the first time, you might infer that a new bike route has just opened up. This process of *action interpretation* can be cast as inference about the world having observed the agent's interactions with the environment (Baker et al., 2009). Formally, this corresponds to Bayesian belief-updating:

$$b'(W, G_I \mid D) \propto P(D \mid W, G_I) b(W, G_I). \qquad (2)$$

We can then define an observer model $O_L(b' \mid D, b)$ that represents an agent observing D and updating their beliefs from b to b' according to Equation 2. To highlight the connection to the interpretation of linguistic utterances, we call this a *literal observer* engaged in literal action interpretation.

Belief-directed planning and pragmatic action interpretation

In a communicative demonstration, the demonstrator has not only instrumental intentions, but also belief-directed ones. To return to our example, the fisherman wants not only to catch a fish (an instrumental intention) but to show an observer how to fish (a belief-directed intention).

Just as instrumental planning involves evaluating and choosing plans over a model of the environment, we propose that belief-directed planning involves doing so over a model of the other person's inferences. This means that rather than just planning a sequence of possible actions and instrumental consequences, a demonstrator also plans over how her actions affect observer beliefs.

Specifically, we posit that it involves planning over the rational inferences that an observer would draw by imputing instrumental goals—those specified by Equation 2. In other words, a demonstrator evaluates whether her actions would accomplish her belief-directed goals by asking what inferences an observer would draw from her actions. She models those inferences by assuming that the observer applies Equation 2, modeling her as an instrumental agent. In this way belief-directed actions are "grounded" in the semantics of instrumental goals. (Later we discuss the possibility of more complex planning over still higher-order inferences).

In order to capture this idea formally we distinguish between instrumental and belief-directed goals and planning (Figure 2C). Recall that G_I denotes instrumental goals expressed as utilities; we denote belief-directed goals as G_B. Given a planning model that includes both the environment (W) and the observer (O_L) as well as instrumental and belief-directed goals (G_I, G_B), belief-directed plans, π_B, that "solve" this planning problem are well defined (although we consider the question of computational tractability in the discussion). When a demonstrator enacts this belief-directed plan in the environment, they produce a demonstration D:

$$\begin{aligned} P(D \mid W, O_L, G_I, G_B) = \\ \sum_{\pi_B} P(D \mid \pi_B, W, O_L) P(\pi_B \mid W, O_L, G_I, G_B) \end{aligned} \tag{3}$$

Again, we emphasize that belief-directed planning depends on reasoning about both the world (W) and a literal observer's belief dynamics (O_L). This is because whereas instrumental planning aims to influence aspects of the environment, belief-directed planning aims to cause both environmental and mental effects. In particular, a belief-directed demonstrator plans and acts by reasoning about how her actions influence observer beliefs via their capacity for action interpretation.

Finally, just as an observer can interpret actions in terms of instrumental intentions, he can also interpret them in terms of belief-directed intentions (Figure 2D). For instance, an observer might assume that the demonstrator is attempting to choose maximally informative actions (in order to accomplish her belief-directed goals) and interpret actions in light of this fact. Formally, we can define *pragmatic action interpretation* as recursive reasoning about the world W, literal observer O_L, instrumental goals G_I, and belief-directed goals G_B:

$$\begin{aligned} b'(W, G_I, G_B \mid D, O_L) \propto \\ P(D \mid W, O_L, G_I, G_B) b(W, G_I, G_B) \end{aligned} \tag{4}$$

Analogously with the literal observer, O_L, we can define a *pragmatic observer*, $O_P(b' \mid D, b)$, who updates their beliefs according to Equation 4. Figure 2 visualizes the functional relationships between the different model components as Bayesian Network Diagrams (Pearl, 1988) from the perspective of the different demonstrators and observers.

Higher-order and mixed-order planning and action interpretation

Equations 1-4 define a sequence of recursive planning and inference processes. In theory, one could have observers who reason about qualitatively different demonstrators as well as demonstrators who reason about more sophisticated observers. For instance, an observer could be *uncertain* about whether they should interpret actions pragmatically: They could be reasoning jointly about a demonstrator's instrumental beliefs (e.g., where do they think the cookies are?) as well as whether they have belief-directed intentions (e.g., are they trying to show me where the cookies are?). Similar situations have been studied in the context of epistemic trust (Mascaro & Sperber, 2009; Shafto, Eaves, Navarro, & Perfors, 2012), where knowledge and helpfulness are uncertain. We could also consider a demonstrator who wants to show an uncertain but pragmatic reasoner that they have a belief-directed intention (e.g., signal that they are signaling; Scott-Phillips et al., 2009). These forms of inference and planning can be formalized in terms of higher-order and mixed-order observers and demonstrators.

In this paper, our primary goal is to understand the first step of the process that relates goal-directed action to communicative action, so we largely focus on straightforward instances of belief-directed planning and pragmatic action interpretation. Nonetheless, we touch on questions about more complex reasoning throughout the paper and return to them in more detail in the general discussion.

Implementations of belief-space planning

Recursive mentalizing and model-based planning are both computationally intensive, as is their combination. The work presented here is not committed to a specific cognitive strategy that people use to compute near-optimal solutions to belief-state planning problems. Rather, our aim is to present a computational-level account (Marr, 1982; Anderson, 1990) that characterizes the problem that people are solving when generating and interpreting communicative demonstrations.

Nonetheless, to generate predictions and evaluate human data requires a specific implementation of belief-space planning (and inference). Briefly, for our Gridworld simulations and analyses, our approach involves first constructing an approximate, discrete belief-space dynamics model that is independent of the particular rewards on a task or trial. This approximate belief-dynamics model is designed to only capture parts of the belief-space that are likely to be visited given an initial belief and environmental dynamics. For a specific trial, this approximate model is combined with an instrumental and/or belief-directed utility function and solved *exactly* using dynamic programming (Bellman, 1957). We note that in contrast to approaches that use sampling to do approximate planning (e.g., Hula, Montague, & Dayan, 2015), this implementation allows us to straightforwardly compute expected rewards and action probabilities that can be used for analysing human responses as well as defining higher-order observers.

Further details about our implementation are reported in the appendix and code itself is available at `https://github.com/markkho/comdem-data-code`. Finally, although we largely sidestep issues of computational cost here, we will return to these questions in our discussion of future work.

Experimental Studies of Communicative Demonstration

Our account provides both quantitative and qualitative predictions about demonstrator actions and observer inferences. We test these in a Gridworld paradigm in which participants played the role of demonstrator that could move a circle on a grid of colored tiles, or observer who was shown a demonstrator's behavior. Experiments 1a and 1b focus on communication of reward structure, while Experiments 2a and 2b focus on learning relevant causal knowledge. Both sets of studies use goal-directed behavior (i.e., doing an activity) as a baseline to compare communicative behavior (i.e., showing an aspect of an activity). Using a combination of behavioral measures, simulations, and model-fitting, we find that belief-directed planning captures key aspects of people's communicative demonstrations.

Experiment 1: Communicating Reward Structure

Communicative demonstrations can be used to convey several kinds of useful information. One important kind concerns the "reward function"—i.e., information about what is desirable and undesirable in the world. This is often expressed as a relationship between object features and rewards. For example, eating red tomatoes might keep you healthy, while eating green tomatoes makes you sick. Thus, a knowledgeable demonstrator will eat red tomatoes and avoid green ones, and an uninformed observer can infer the true reward structure by observing this.

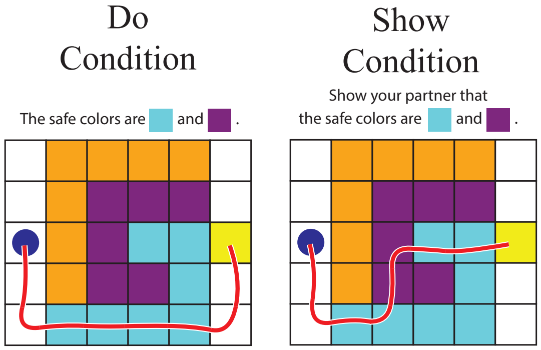
Do Condition
The safe colors are 🟦 and 🟪 .
Show Condition
Show your partner that the safe colors are 🟦 and 🟪 .

Figure 3. Experiment 1 - Participants were either placed in a condition where they were simply told the reward function (left) or also told to show the reward function to a partner (right). The red lines are representative examples of behavior in the two conditions.

Experiment 1 uses this case study of teaching a reward function to test our account of communicative demonstration. Critically, we note the methodological importance of using feature-based rewards: Distinguishing between "doing" an activity and "showing" relies on the possibility of information that generalizes beyond the immediate activity. This is because if a demonstrator is only showing how to do the immediate activity, the best strategy is to simply do the activity and have the learner copy those exact actions. Moreover, as we explore in the discussion, this experimental consideration is closely related to theoretical claims that relate communicative demonstrations to the transmission of generalizable knowledge (Csibra & Gergely, 2009).

Thus, in Experiment 1a, we focus on how people in the role of demonstrator convey feature-based—that is, generalizable—information about rewards. Using a combination of behavioral and model-based analyses, we show that people motivated to demonstrate task information uniquely engage in strategies that reflect a combination of instrumental and belief-directed planning. For example, we expect people to engage in "targeted variability" where they strategically visit tiles that have high diagnostic value. We first look for these types of behavioral signatures predicted by the model before reporting the results of model-fitting and parameter estimation. Experiment 1b then focuses on observer judgments. We find that people make more accurate and confident inferences when demonstrations are known to be communicatively generated, consistent with our model of pragmatic action interpretation.

Method

Task. Participants were asked to navigate the Gridworld shown in Figure 3 by moving the blue circle up, down, left, or right on each time step. Yellow tiles were always "goal" states, meaning that whenever it was entered, the participant received +10 points and the trial ended. White tiles always reward 0 points. The remaining tile colors—orange, purple, and blue—each were randomly assigned to be either *safe* or *dangerous*, meaning that their reward values could either be 0 or -2 points respectively. Thus, the set of possible reward structures formed by all combinations of safe or dangerous yields a space of eight possible distinct reward structures.

Procedure. Sixty participants recruited from Amazon Mechanical Turk performed the feature-based reward teaching task; two were excluded due to missing data due to recording error, leaving a total of 58 participants for analysis (29 in each condition). They received a base pay of $1.00 and received a bonus based on points received across the whole experiment, with each point worth +/- 2 cents. The experiment was organized into a training phase and test phase. The training phase was designed to familiarize participants with the domain by alternating between learning a reward function and then applying it. On the learning trials, they were not told the underlying reward structure (i.e. which colors were safe/dangerous), but received immediate feedback on how many points were won or lost when stepping on tiles. On the applying trials immediately following each learning trial, they were given a new grid configuration that required knowledge of tile color type, and applied what they just learned about the tiles without receiving feedback. They repeated this procedure 8 times for each of the 8 possible combination of "safe" and "dangerous" colors. The order of the reward functions was randomized between participants.

Following the training phase participants were split into two conditions: One that only motivated completing the task (the "Do" condition) and another that additionally motivated demonstrating the task to an observer (the "Show" condition). Both Do and Show participants were told which colors were safe and won or lost points based on which tiles were safe or dangerous. Only Show participants were additionally told that their behavior would be shown to another person, that this person could not see the points they received, that they would apply what they learned to a new grid, and that the points won by their partner would be added to their bonus. When bonuses were calculated, participants each received what they would have had their partner done as well as possible. Participants did not receive feedback on the reward for each action, although this could be easily inferred from the information provided. Procedures were approved by Brown University's Research Protection Office (protocol #1505001248, title: "Exploring human and machine decision-making in multi-agent environments").

Simulated Demonstrators

We generated simulated behaviors of instrumental and belief-directed demonstrators from our model for each of the eight reward structures. Specifically, for each of the eight trials, 200 sequences of states and actions were generated. Half of these were from the instrumental demonstrator, which correspond to the Do condition, while the other half were from the belief-directed demonstrator, which correspond to the Show condition. Details for how we formalized the task and models are described in the supplementary materials.

Results

The open-ended nature of the task led to a range of participant demonstrations, visualized in Figure 4a. These data largely matched the qualitative and quantitative predictions of the model simulations: Do participants selected routes based exclusively on efficiency, whereas Show participants took routes that additionally signaled feature reward values. To understand people's behavior in light of the model, we performed three sets of analyses. First, we examined task-specific behavioral predictions based on our simulated demonstrations. Specifically, we examined the number of color tiles and the variability of color tiles visited as a signature of belief-directed planning. Second, we examined how the sequences of actions people took in each condition led to transitions in the belief-space of several observer models. Finally, we performed a model comparison analysis to confirm that the behavior in Show is explained by belief-directed planning and not a particular parameterization of pure instrumental planning.

Behavioral Analysis. Our model predicts that instrumental action and communicative demonstrations will differ from one another in systematic ways. For instance, on trials where multiple colors are safe (e.g., orange and blue in Figure 3, it may be worthwhile for belief-directed planning to engage in "targeted variability" where multiple tiles types are visited, whereas pure instrumental planning would lead to visiting only one or the other if it is maximally efficient for reaching the goal. We quantified these types of predicted differences by calculating the proportion of orange, blue, or purple tiles visited in a trajectory (*color visitation proportion*) and the entropy of the frequency distribution over orange, blue, and purple tiles in a trajectory (*color visitation entropy*). As shown in Figure 5A (top row), both color visitation proportion and entropy was generally higher for the belief-directed model, although this varied by the particular reward function.

We then calculated color visitation proportion and entropy for the empirical trajectories and performed two sets of analyses. First, we analyzed the trajectories independently of the planning models using a mixed-effects logistic regression for color visitation proportion and a mixed-effects linear regression for color visitation entropy. For both of these models, condition was included as a fixed effect while by-participant and by-item (i.e., reward function) random intercepts were fit. Show condition trajectories had a greater color visitation proportion ($\beta = 2.59$, $SE = 0.47$, $Z = 5.46$, $p < .0001$ [Wald Z test]) as well as color visitation entropy ($\beta = 0.16$,

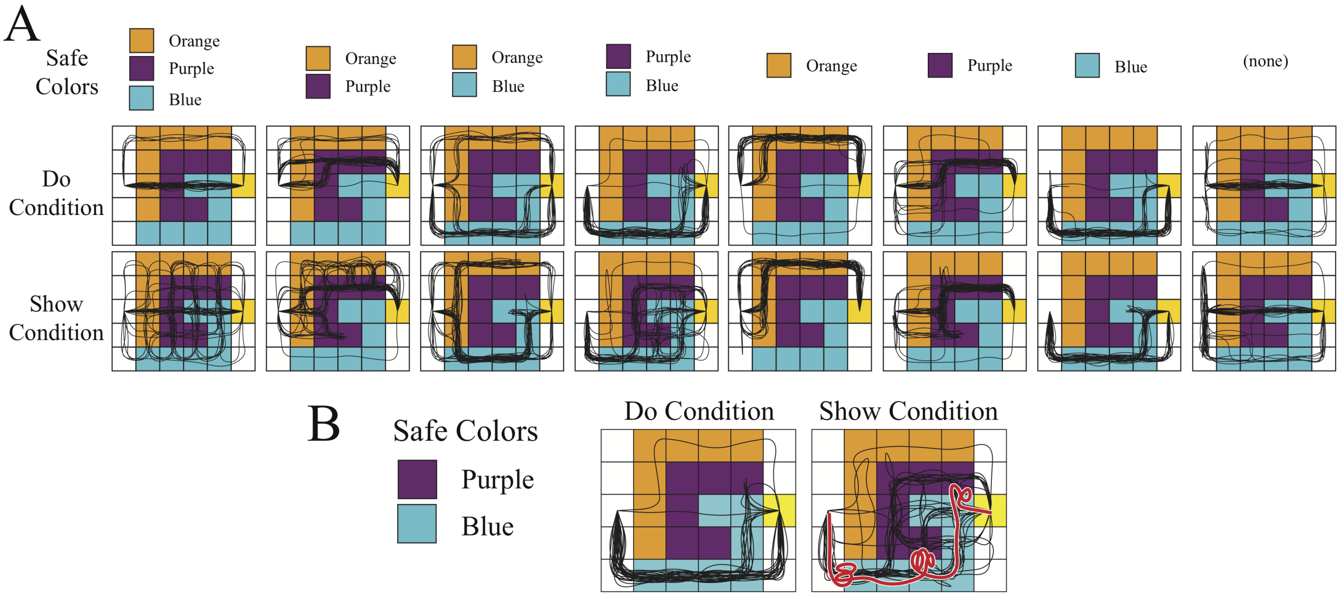

Figure 4. Experiment 1 results. (a) Participant demonstrations by trial. Each line is one participant's demonstration. Column labels indicate codes for each reward structure. Top row: Ground rewards of the task. Yellow goal tiles are worth +10 points, colors that are safe are 0 points, otherwise they are −2 points. Middle row: Do participant trajectories with the tile colors that were visible to the participants (orange, purple, blue, white and yellow); Bottom row: Show participant trajectories. (b) An exaggerated demonstration (red line) in the Show condition. Behaviors such as this are predicted by the planning in observer belief space model. Panel (a) adapted from Ho et al. (2016).

$SE = 0.05$, $t(57.04) = 3.30$, $p < .01$ [Satterthwaite's approximation]).

Second, we regressed the color visitation measures from participant trajectories onto the mean simulated trajectory values using logistic and linear mixed-effects models, respectively. For these models, we included the simulation predictions as a fixed effect and by-participant and by-item intercepts as random effects. We found that the measures produced by the instrumental and belief-directed model simulations were predictive of both color visitation proportion ($\beta = 2.59$, $SE = 0.47$, $Z = 5.46$, $p < .0001$ [Wald Z test]) and entropy ($\beta = 0.92$, $SE = 0.06$, $t(9.41) = 14.37$, $p < .001$ [Satterthwaite's approximation]).

Collectively, these sets of analyses indicate that qualitative differences related to targeted variability between the Do and Show conditions matched those of the instrumental and belief-directed planning models.

Belief-Space Analysis. The behavioral analyses provide insight into people's demonstrations in terms of features of the ground task itself (e.g., what colors were visited). To gain initial insight into the effects of people's demonstrations on belief-dynamics, we analyzed belief-space transitions for three types of observer models. To be clear, our goal is not to compare observer models, but to characterize the inferential effects of people's behavior on several rational observer models.

First, we examined the effect on a literal observer. For each participant trajectory, the model's final belief in the true trial parameters was used as a predictor in a mixed-effects linear model that included by-participant and by-reward function random intercepts as well as condition (Do/Show) as a fixed effect. Condition was found to be significant ($\beta = 0.19$, $SE = 0.03$, $t(56.00) = 5.54$, $p < .001$ [Satterthwaite's approximation]), indicating that trajectories in Show led the observer model to have a higher belief on the target.

Second, we compared how a pragmatic observer who reasons about the demonstrator's actions as intentionally informative would learn. Using the same analysis with final beliefs as the predictor in a mixed-effects linear model, we found Condition was similarly significant ($\beta = 0.23$, $SE = 0.04$, $t(56.00) = 5.74$, $p < .001$ [Satterthwaite's approximation]), confirming that a pragmatic observer would also learn better from these trajectories.

A literal observer assumes the demonstrator definitely does not have an intention to inform, while the pragmatic observer assumes the complete opposite. But what if the observer is uncertain about the demonstrator's communicative intent? Analogous situations have been studied in the context of epistemic trust (Mascaro & Sperber, 2009) and can be modeled as joint inference over another's knowledge and helpfulness (Eaves Jr & Shafto, 2012; Eaves & Shafto, 2017; Shafto, Eaves, et al., 2012). To test how such an observer would respond to people's demonstrations, we implemented an *uncertain pragmatic observer*, who reasons jointly about the task parameters and whether the demonstrator has informative intentions. Using the same analysis over final belief in the trial parameters as before, we find that condition was significant ($\beta = 0.23$, $SE = 0.03$, $t(56.00) = 6.45$, $p < .001$ [Satterthwaite's approximation]). Additionally, we used a similarly structured analysis to determine if the uncertain pragmatic observer tracks whether the demonstrator was attempting to be informative. We found a significant effect of condition ($\beta = 0.09$, $SE = 0.03$, $t(56.00) = 3.51$, $p < .001$ [Satterthwaite's approximation]), which confirms that people's actions in Show are not only more informative, but interpreted as more intentionally informative by an uncertain model.

In summary, we examined how people's behaviors in Do and Show affected the beliefs of three types of observer models (literal, pragmatic, and uncertain pragmatic). People's behavior in Show better conveyed the underlying structure of the task for all three observer models. Additionally, we found that Show demonstrations are themselves interpretable as intentionally informative to an observer who is unsure.

Model-fitting Analysis. As a final, stronger test of the distinction between doing an activity and showing an activity, we employ model fitting. This approach ensures that the effectiveness of Show trajectories is due to belief-directed planning and not a particular parameterization of instrumental planning. For example, people could simply act more randomly in the Show condition rather than engage in targeted variability. Additionally, model-fitting allows us to understand people's low-level actions—e.g., movements in the cardinal directions—in terms of high-level psychological and computational constructs—e.g., communicative utilities and recursive models of an observer. Here, we describe our general results and report details related to implementation and parameter estimates in the supplementary materials.

Our main question is whether including belief transitions and utilities in a demonstrator's planning model explains behavior in Show but not in Do. To assess this, we fit maximum-likelihood parameters for the instrumental planner and for the belief-directed planner to each participant. Since instrumental planning is a nested version of belief-directed planning (i.e., where belief transitions and utilities are ignored), we compared model-fits in each condition using a log-likelihood ratio test with three degrees of freedom difference per participant. For Do, the instrumental model was not rejected ($\chi^2(87) = 100.78$, $p = .15$), whereas for Show, it was ($\chi^2(87) = 391.60$, $p < 10^{-39}$), indicating that the belief-directed planning model provides a better explanation of behavior for the Show but not Do. Additionally, we can compare models for individual participants. Figure 5c shows the likelihood-ratio test statistic associated with each participant. We used a permutation test (Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005) with $10,000$ random permutations to determine whether the difference in mean test statistics between Do and Show was significant (Do *LR* mean = 3.47, S.D. = 4.63; Show *LR* mean = 13.50, S.D. = 8.78). None of the permutations exceeded the true difference in mean test statistic ($p = \frac{1}{10001} < 10^{-4}$) indicating that more participants in Show are explained by belief-directed planning.

Given belief-directed planning models fit to each individual participant's collection of demonstrations, we can also examine the parameters corresponding to planning goals and representations. In particular, we examined two parameters corresponding to the strength of the demonstrator's communicative goal (G_B) and the informativeness of instrumental actions for a literal observer (O_L) in light of the Do and Show conditions. As expected, we find that estimated communicative goal strength was higher for Show than Do (Wilcoxon Signed-Ranks test: $Z = -2.11, p < .05$). Similarly, we find that instrumental action informativeness to be higher for Show than Do (Wilcoxon Signed-Ranks test: $Z = 1.98, p < .05$). Moreover, these two parameter estimates are correlated in Show but not in Do, indicating a coupling between the representational and motivational dimensions of communicative demonstrations (Do: $r = .27, p = 0.16$; Show: $r = .50, p < .01$). These analyses of individually fit parameters provide further confirmation that the behavior in Show results from a planning process informed by literal action interpretation. For complete details on how these parameters were specified, please see the supplementary materials.

To summarize, we find that belief-directed planning provides a better model of behavior in Show than in Do.

Experiment 1b: Learning Reward Structure from Demonstrations

We next turn to observer behavior. Specifically, we evaluated whether communicative demonstrations are effective for teaching human observers the true reward function, and also whether observers' expectations of communicative ("showing") or non-communicative ("doing") behavior matter. Participants were either placed in a Communicative or Non-Communicative interpretation condition, corresponding to the literal and pragmatic observer models, respectively. They were then given the trajectories from either the Show or Do conditions in Experiment 1a. Overall, we find a large positive effect on observer accuracy and confidence when given demonstrations from Show versus Do, and a small positive effect of observer interpretation consistent with the model predictions.

Materials and Procedure. The stimuli were the state/action/next-state sequences produced by participants in Experiment 1a. These were generated from the eight critical trials from the 29 participants the Do/Show demonstrator conditions, for a total of 464 demonstrations. Each participant was told they would observe a single demonstration from a partner. They were also assigned to a Communicative or Non-Communicative interpretation condition. The in-

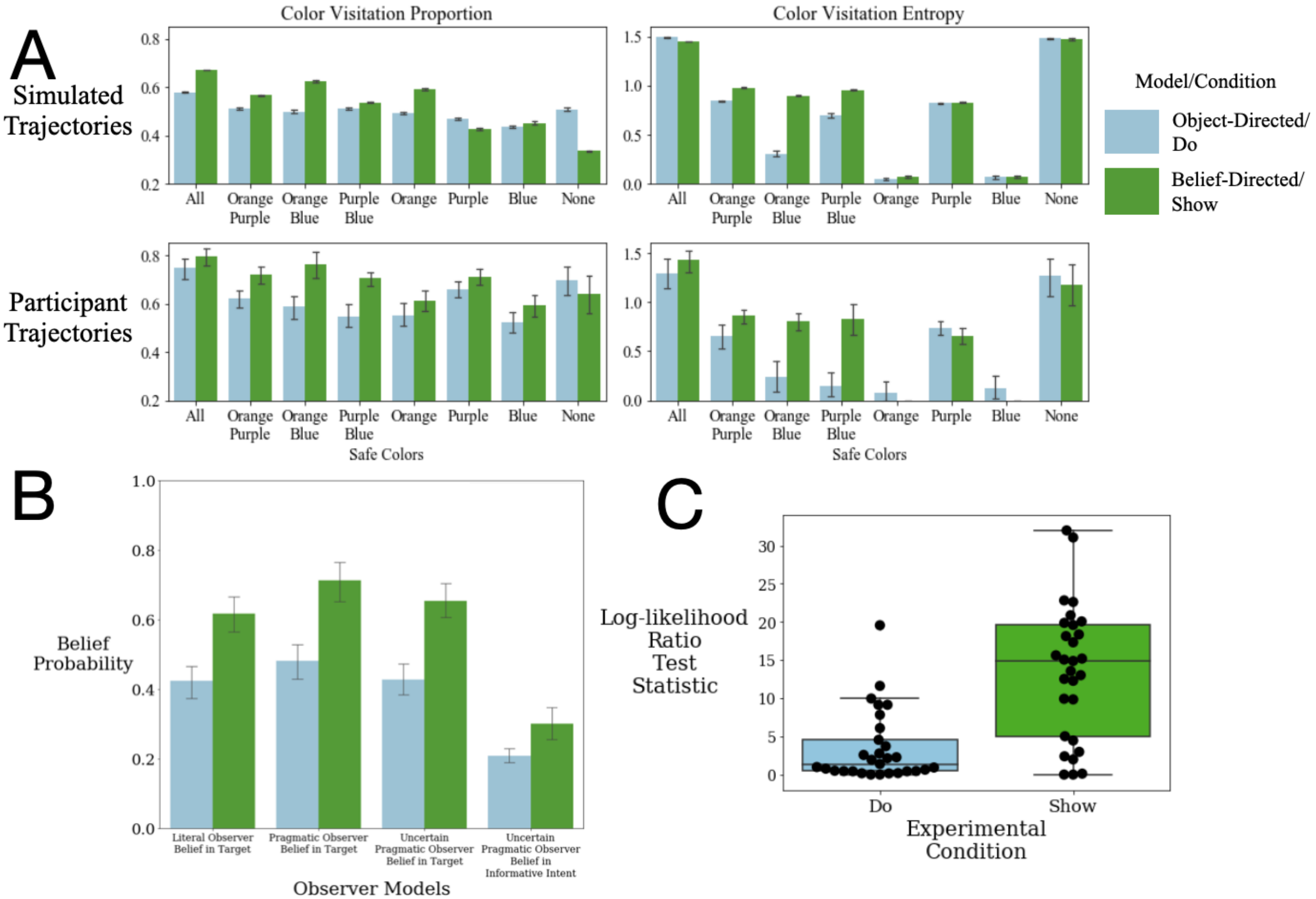

Figure 5. Experiment 1a Analyses. (A) Comparison of color visitation proportion and color visitation entropy for simulated trajectories (top row) and participant trajectories (bottom row). Each pair of bars corresponds to one of the 8 reward functions. Analyses with mixed-effects linear models revealed a significant effect by Do/Show condition and that simulation values predict experimental measures. Error-bars are bootstrapped 95% confidence intervals. (B) Belief-space analysis. We provided participant-generated trajectories from the Do and Show conditions to three different observer models and measured final beliefs. A literal observer updates his beliefs about the world assuming an instrumental demonstrator, while a pragmatic observer updates his beliefs assuming the demonstrator is belief-directed and attempting to be informative. The uncertain pragmatic observer has probabilistic beliefs about both the world as well as whether the demonstrator is belief-directed. Empirical Show trajectories are consistently better at conveying task structure across all three observer models and are interpreted by the uncertain pragmatic observer model as being more intentionally informative. (C) By-participant model fitting results. Each point represents the log-likelihood ratio test statistic comparing the instrumental planning model to the belief-directed planning model. Belief-directed planning better accounts for demonstrator behavior more in Show than in Do (permutation test, $p < 10^{-4}$; see main text and supplementary materials for details).

structions were the same except participants in the Communicative condition were also told that their partner "knows that you are watching and is trying to show you which colors are safe and dangerous". Next, they were shown a page with the animated demonstration and answered, for each of the three colors (orange, purple, and blue), whether they thought it was safe or dangerous and their confidence on a continuous scale (0 to 100). Each participant received a starting payment of 25¢ and won/lost 5¢ for each correct/incorrect answer (minimum payment was 10¢). Two MTurkers were assigned to each demonstration and observer instruction combination using psiTurk (Gureckis et al., 2016). Procedures were approved by University of Wisconsin-Madison ED/SBS IRB (Study #2017-0830, title: "Studying human and machine interactions").

Simulated Observers. In order to compare model predictions with human performance, we simulated observer beliefs using the literal and pragmatic observer models reported in the belief-space analysis in Experiment 1a. Since participants gave judgments for each color separately, we compared the marginalized probabilities for each color to judgments. Mean correct belief probabilities by observer model and demonstrator condition are shown in Figure 6.

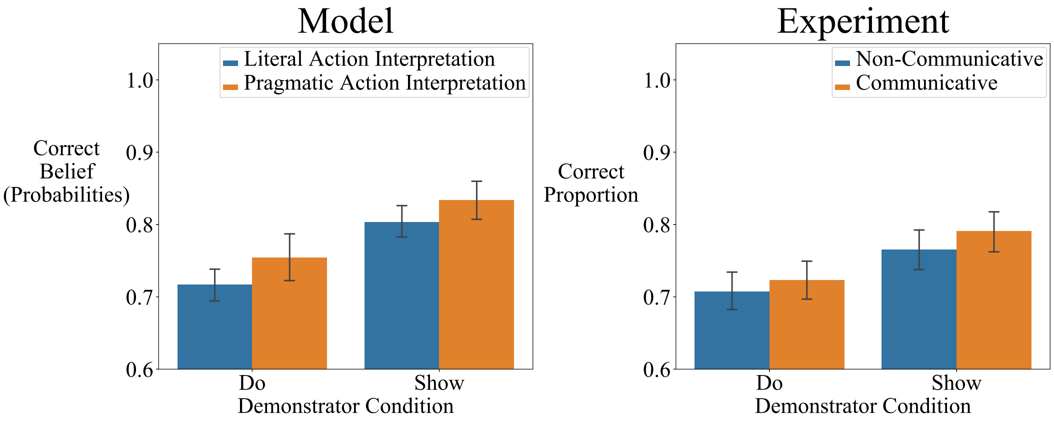

Figure 6. Comparison of model and human observers learning from human demonstrations. Observer models were either literal observers or pragmatic observers, and were given Experiment 1a trajectories from the Do or Show condition. Left panel shows the correct probability that a color is safe (not the probability of the task) in order to facilitate comparison with Experiment 1b. Participants in Experiment 1b were either told the demonstrator was intentionally communicating (Communicative condition) or nothing (Non-Communicative condition), were given Do or Show trajectories, and gave safe/dangerous judgments for each color that were coded as correct or incorrect. Error bars are bootstrapped 95% confidence intervals.

Results. Participants exposed to Show trajectories were more accurate and confident in their beliefs, compared with participants exposed to Do trajectories. To analyze accuracy, we used a mixed-effects logistic regression with correct/incorrect judgments as the outcome variable. By-trial (reward function), by-demonstrator, and by-observer intercepts were used as random effects, and both sets of instructions and their interaction were set as fixed effects using contrast coding. The effect of whether the trajectories were from Do or Show (demonstrator instructions) was significant ($\beta = 0.40$, $SE = 0.11$, $Z = 3.64$, $p < .001$ [Wald Z test]) as was the effect of the Communicative/Non-Communicative interpretation (observer instructions) ($\beta = 0.13$, $SE = 0.07$, $Z = 2.02$, $p < .05$ [Wald Z test]). Demonstrator Show instructions had a larger effect size, corresponding to an increase in observer accuracy by 1.5 times, as compared to observer Communicative instructions, which corresponds to an increase in observer accuracy by 1.14 times. There was no significant interaction ($\beta = 0.09$, $SE = 0.14$, $Z = 0.64$, $p = .52$ [Wald Z test]). This general pattern parallels the simulation results (Figure 6).

Confidence judgments were analyzed by mixed-effects linear regression. Reported confidence was the outcome variable; trial, demonstrator, and observer were random effects; and demonstrator instructions, observer instructions, and their interaction were fixed effects. Observers receiving Show demonstrations were more confident ($\beta = 3.34$, $SE = 0.93$, $t(57.20) = 3.59$, $p < 0.001$ [Satterthwaite's approximation]), as were those receiving Communicative instructions ($\beta = 3.57$, $SE = 0.87$, $t(1790.85) = 4.08$, $p < 0.0001$ [Satterthwaite's approximation]). There was no significant interaction ($\beta = 1.16$, $SE = 1.75$, $t(1790.85) = 0.67$, $p = .51$ [Satterthwaite's approximation]). In short, the generation and interpretation of demonstrations as communicative increased both accuracy and confidence.

Discussion

Are people's communicative demonstrations reflective of belief-directed planning? Experiment 1 addressed this question in the context of communicating information about reward structure. To test this, in Experiment 1a participants performed a series of tasks that differed by reward structure and were motivated to merely "Do" or additionally "Show" the task. In Experiment 1b, these demonstrations were observed by a separate set of participants, who were either told that the demonstrations were produced communicatively or not told this. These participants were then asked for their beliefs about the reward structure.

Our analyses of demonstrators' behavior as well as observers' judgments provide evidence for belief-directed planning and pragmatic action interpretation described by our account. First, we show that the belief-directed demonstrator model predicts how people will modify their goal-directed behavior by increasing the variability of color tiles visited on specific trials. Second, using three different observer model variants, we find that Show demonstrations more effectively convey the underlying task structure and that they are more likely to be interpreted as intentionally informative by a rational model. Additionally, the qualitative pattern of judgments provided by two of these models (the literal and pragmatic observer) match those of participants in Experiment 1b. Finally, the model-comparison analysis of Do/Show demonstrations allows us to conclude that the variability in actions taken in communicative demonstrations are not simply due to undirected noise, but rather belief-directed planning.

In short, Experiment 1 tests the predictions of our demonstrator and observer models in the context of conveying information about reward structure. Overall, we find that the differences between planning and interpreting instrumental actions versus communicative demonstrations can be understood in terms of our account of recursive planning and inference.

Experiment 2: Communicating Causal Structure by Demonstration

Actions can convey more than just the reward structure of the world; they can also convey its causal structure. For instance, consider showing someone else how to use a can opener. You would want to make sure that they learn the key causal dependencies between squeezing the handles and puncturing the lid, and between rotating the handle and cutting the lid free. Exaggerating certain motions involved in using a can opener can provide clear evidence of the underlying causal mechanism and would directly result from belief-directed planning, even though these exaggerations may not be the most efficient way to actually open the can.

In Experiment 2, we examine how people modify their instrumental actions to convey information about hidden causal structure. We used a variation on the Gridworld task in which certain tiles have different probabilistic outcomes (Figure 7), allowing us to examine how people exploit complex dynamics of an environment when acting communicatively. In particular, having probabilistic causal affordances enables us to directly test whether communicative demonstrators are leveraging observers' capacity for theory of mind when planning over beliefs. This is because such situations allow demonstrators to show that they are *trying* to do something, even if it is potentially costly or may not succeed. For instance, suppose you want to show someone how to use a can opener, but it fails on some cans and breaks them because it is of poor quality. You may still be able to convey how can openers work by using properly exaggerated motions to indicate how you *expect* it to work, even if it does not *actually* work the way it is supposed to. In this experiment, we introduce "jumper tiles" that allow agents to jump over dangerous tiles. This provides opportunities for participants to show observers when certain tiles are jumpers by taking extra jumping actions as well as showing them that they can be used to avoid dangerous tiles by taking *risky jumps* that sometimes do not work.

As we discuss, our framework for planning and interpretation of communicative demonstrations allows us to compare belief-directed planning not only to instrumental planning, but also to variants that make weaker assumptions about observers. For instance, whether an observer engages in inverse planning or simple causal reasoning will affect whether a demonstrator engages in risky jumping. Overall, we find that the full belief-directed planning model uniquely predicts how people will act to convey causal structure through their actions.

Method

Task. We used the layout shown in Figure 7a to test how people convey causal structure. Each trial, people start at the bottom center of the grid and must reach the yellow goal tile (worth 50 points) in as few steps as possible (each step was penalized -1 point). Dangerous tiles (red) are always worth -10 points. "Jumper tiles" (green) are worth zero points, but stochastically cause the agent to jump over the immediate tile, thus avoiding losing points if it is dangerous. Within a trial, all of the jumper tiles are either "strong", meaning that 3/4ths of the time the tile moved the agent two steps and 1/4th of the time moved it only one step, or "weak", in which the probabilities were reversed. As a result, the value of actions from a particular jumper tile depends on both the layout of the dangerous and jumper tiles, as well as whether the jumper tiles are strong or weak.

Procedure. 80 Amazon Mechanical Turk participants participated for payment. The overall design of this experiment was similar to that of Experiment 1 with a few modifications. Participants were trained on the basic experimental interface and interacted with a set of 16 exploration grids in which they had to figure out whether the jumper tiles were strong or weak. The grids were designed such that there was no way to try a jumper tile without some risk of entering a dangerous tile. After each of these exploration rounds, they had to answer whether they thought the jumper tiles on that trial were strong or weak and won or lost 50 points based on their answer. They were then split into two conditions: Do and Show. Forty-one participants were assigned to Show while 39 were assigned to Do. In Do, participants were always told whether the jumper tiles were strong or weak; in Show, they were also told this information but were additionally told that their behavior would be shown to a partner who would have to answer whether the jumper tiles were strong or weak. They would then win or lose 50 points based on their partner's answer.

Both conditions were given the same set of 8 grids twice. We designed the grids to favor certain trajectories when jumper tiles were weak (*weak affording*), others when jumper tiles strong (*strong affording*), and others whether the jumper tiles were weak or strong (Figure 7b). We arranged these variations to distinguish between doing and showing, because sometimes the most effective way to communicate task structure would be to incur the risk of jumping onto a dangerous tile. Each grid was then presented where the jumper tiles were strong and weak, for a total of 16 distinct rounds per person. Procedures were approved by University of Wisconsin-Madison ED/SBS IRB (Study #2017-0830, title: "Studying human and machine interactions").

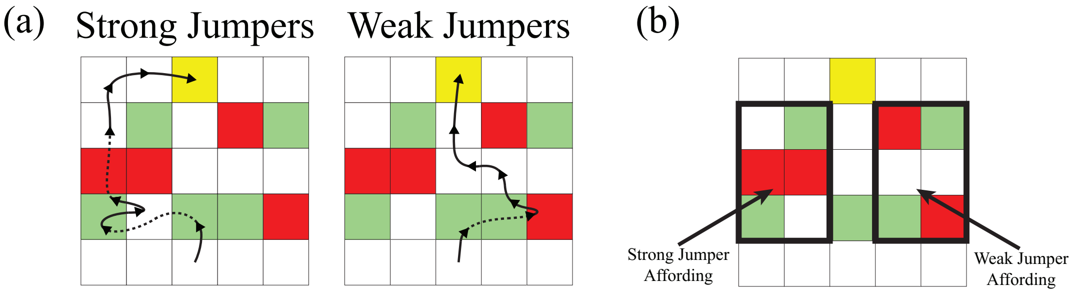

Figure 7. Experiment 2 paradigm. Each trial had a particular configuration of dangerous tiles (red) and jumper tiles (green). (a) Example trajectories showing whether the jumper tiles are *strong* (the agent is usually moved two tiles after stepping off it) or *weak* (agent is only sometimes moved two tiles). Dotted line indicates a successful jump of two tiles. Within a trial, jumper tiles are either all strong or all weak. (b) Example of regions of grid with different affordances based on whether the jumper tiles are strong or weak.

Simulated Demonstrators

Similar to Experiment 1, we simulated both instrumental and belief-directed demonstrators for both strong and weak variants of the Gridworld trials. However, we also generated two types of belief-directed demonstrator. The first was a demonstrator who planned over a literal observer who performed inverse planning—i.e., it assumes that the observer reasons about instrumental intentions. The second was a demonstrator who planned over a *lesioned* literal observer that does not perform inverse planning and can only do causal reasoning. Fifty sequences of states and actions were generated for each of these three models for each transition function and Gridworld.

Results

As in Experiment 1a, we performed three sets of analyses to understand participants' behavior (Figure 8): task-specific behavioral predictions, performance in model observers' belief-space, and by-participant model-fitting. For the behavioral analyses, we focused on jumping rate and risky jumping rate, as these are key signatures of belief-directed planning. Risky jumping in particular allows us to contrast demonstrations designed for observers engaging in inverse planning versus one that only does causal reasoning. Overall, these analyses allow us to evaluate whether people's communicative behavior is explained by planning over a model of observers' beliefs from several different perspectives.

Jumping and Risky Jumping. Jumpers allow people to take inconvenient or risky actions to convey information about the task. For example, repeatedly jumping off of green tiles can provide evidence of the underlying causal mechanism by providing observers the opportunity to directly observe the relevant statistics. However, our set up also provides demonstrators with an opportunity to convey their *expectation* that a particular outcome is likely by taking *risky jumps*, where possible outcomes have a large influence on rewards. For instance, if green jumper tiles were strong, an agent could try to use it to jump over a red tile, whereas if it were weak, they would not. If the observer can reason about an actor's intention to use jumpers to skip over red tiles—that is, if they can engage in literal action interpretation—, then a communicative demonstrator could use risky jumping to signal their expectations. In contrast, a rational communicative agent who does not think the observer can reason about intentions would not take risky jumps if they did not have to. These two possible communicative agents correspond directly to belief-directed planning over a full literal observer and planning over a lesioned observer that we simulated. As shown in Figure 8A (top row), although both types of agents engage in more jumping than the instrumental agent baseline, only planning over a full literal observer predicts more risky jumping.

We analyzed participants' jumping and risky jumping independent of the models and then with the model predictions. For our first set of analyses, we fit a linear mixed-effects model to the number of jumps per round and a logistic mixed-effects model to whether jumps were risky. A jump was defined as any action that had a non-zero probability of moving two tiles, while a risky jump was coded as any jump that had a non-zero probability of landing on a red tile. Both models included condition (Do/Show) as a fixed effect as well as by-participant and by-item random intercepts. The Show condition had significantly more jumps per round ($\beta = 0.71$, $SE = 0.14$, $t(79.71) = 5.22$, $p < .0001$ [Satterthwaite's approximation]) and jumps that were risky ($\beta = 0.98$, $SE = 0.24$, $Z = 4.16$, $p < .0001$ [Wald Z test]), matching the qualitative patterns of belief-directed planning.

To assess whether belief-directed planning over the full literal observer predicted risky jumping over and above planning over the lesioned literal observer, we compared two nested logistic regression mixed-effects models. Both models set whether a jump was risky as the dependent variable, included by-participant and by-item random effects, and used data from both Do and Show conditions. The first model contained only the lesioned planning risky-jump proportions as a fixed effect for each grid, transition strucutre (strong/weak),

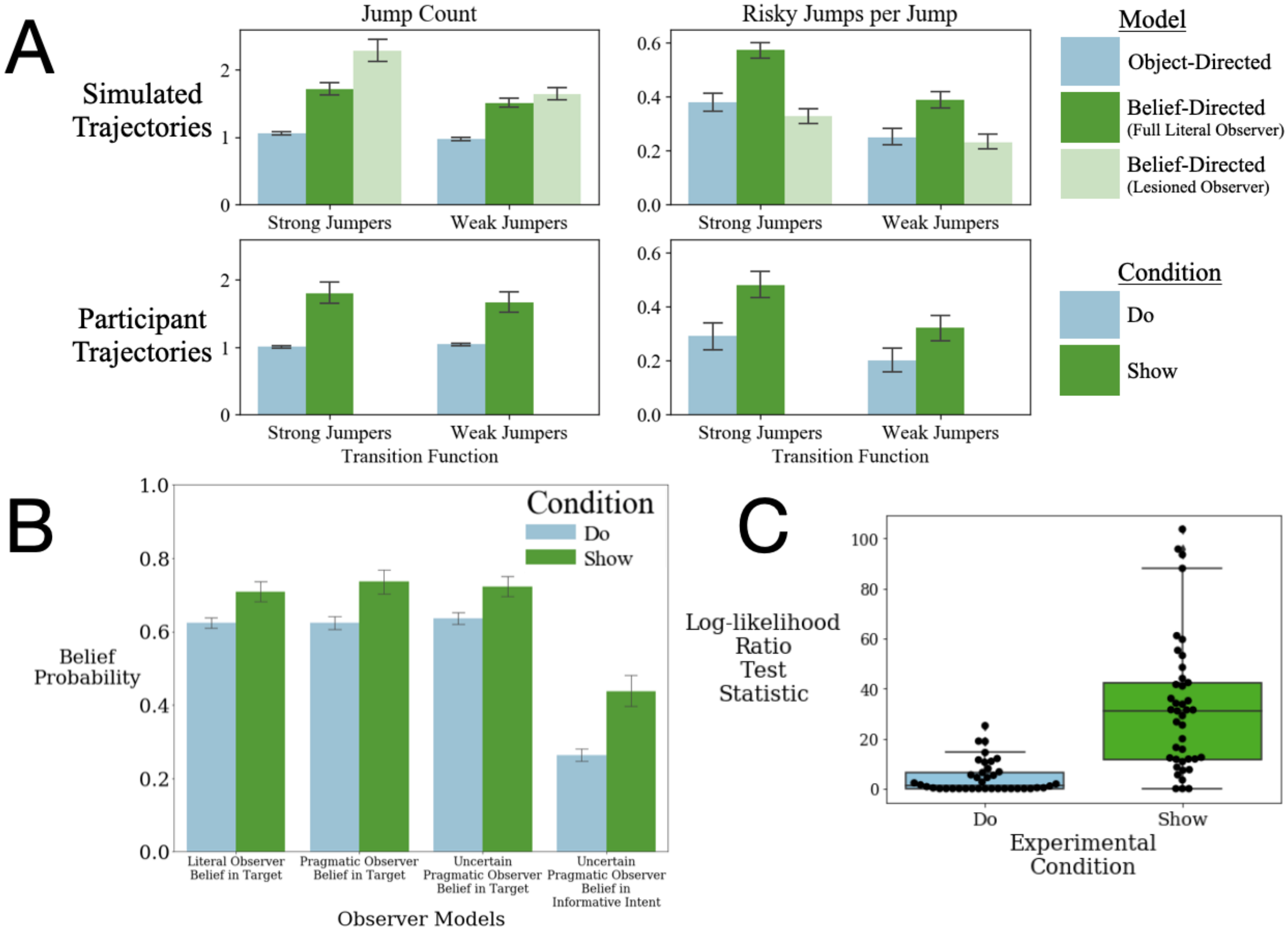

Figure 8. Experiment 2a results. (a) Comparison of simulated (top) and experimental jumping (bottom). We simulated instrumental planning actions (blue bars) as well as two types of belief-directed planning demonstrations (dark and light green bars). The first belief-directed planner calculated a plan over a full literal observer model capable of reasoning about instrumental intentions (e.g., whether the agent was *trying* to jump). The second belief-directed planner was provided with a lesioned literal observer model that could not reason about instrumental intentions and could only draw inferences based on observed outcomes (e.g., whether the agent *actually* jumped). In particular, note that because inference about risky jumping requires that the observer reason about planning (an action is only risky with respect to the agent's intention not to lose points), only the first belief-directed planner will engage in additional risky jumping. (b) Belief-space analysis. Participant trajectories were provided to three different observer models: a literal observer; a pragmatic observer; and an uncertain pragmatic observer. We examined how the trajectories from each participant condition influenced the beliefs of each observer model. (c) Model-fitting Analysis: Each point represents the log-likelihood ratio test statistic comparing the instrumental planning model to the belief-directed planning model. Belief-directed planning better accounts for demonstrator behavior more in Show than in Do (permutation test, $p < 10^{-4}$; see text for details of analysis).

and condition, while the second additionally included the proportions from belief-directed planning over the full literal observer. Note that the predictors for Do condition jumps were always those of a planner for whom the observer belief model is irrelevant, that is, the pure instrumental planner. Using log-likelihood ratio tests, we found that including the full planning predictions produced a significantly better fit ($\chi^2(1) = 12.42, p < .001$), indicating that people's risky jumping is explained by belief-directed planning over a full literal observer performing inverse planning.

In sum, if people are engaging in belief-directed planning over a literal observer model in Experiment 2b, they will be more likely to use jumpers, especially when they are risky, if it effectively signals the underlying causal structure of a task. We find that people do engage in these behaviors in the manner predicted by the full belief-directed planning model.

Belief-space Analysis. We next ask whether actor behavior in the Show condition conveys more information about state transitions than actor behavior in the Do condition. To do this, we assessed changes in the belief space of three types of model observers as in Experiment 1a: a literal observer, a pragmatic observer, and an uncertain pragmatic observer. The observer models were generated using the same procedure as in Experiment 1a.

We used the same mixed-effects model as in the previous analyses to determine whether Show and Do demonstrations resulted in different final beliefs. All three observer models learned the target better from the Show than Do demonstrations (Show/Do fixed effect for Literal Observer belief in strong/weak: $\beta = 0.08$, $SE = 0.02$, $t(78.00) = 5.03$, $p < .001$; Pragmatic Observer: $\beta = 0.11$, $SE = 0.02$, $t(78.00) = 5.82$, $p < .001$; Uncertain Pragmatic Observer: $\beta = 0.09$, $SE = 0.02$, $t(78.00) = 5.21$, $p < .001$ [Satterthwaite's approximation]). Additionally, for the uncertain pragmatic observer, we found that the model interpreted Show demonstrations as being more likely to be communicative (Show/Do fixed effect: $\beta = 0.17$, $SE = 0.02$, $t(78.00) = 7.20$ [Satterthwaite's approximation], $p < 0.001$). Figure 8b plots mean beliefs by condition for each observer model. Collectively, these analyses indicate that participants in Show chose action sequences that are consistently successful in modifying the belief state of an observer towards the target belief.

Model-fitting Analyses. For our final set of analyses, we used by-participant model fitting and model comparison to determine whether belief-directed planning explains behavior in the Show condition but not the Do condition. We performed the same analyses as in Experiment 1a. Likelihood ratio tests at the condition level revealed that both Do and Show behavior was explained better by belief-directed planning (Do: $\chi^2(117) = 174.27$, $p < .0001$; Show: $\chi^2(123) = 1328.24$, $p < 10^{-200}$). However, a comparison of participant-level likelihood-ratio test statistics revealed a clear difference in how well the models account for behavior in each condition (Figure 8C). Specifically, we used a permutation test to compare means of the likelihood ratio test statistic in each condition ($10,000$ random permutations) and found none of the permutations exceeded the true difference in mean test statistic ($p = \frac{1}{10001} < 10^{-4}$). This indicates that belief-directed planning captures variability in Show behavior that is distinct from Do.

Additionally, the participant-level fits allow us to assess individual parameter estimates and whether they reflect the belief-directed planning goals and representations in our model. As in Experiment 1a, we examined parameters associated with communicative goal strength and instrumental action informativeness (full details on how these parameters are specified are available in the supplementary materials). As expected, we found that estimated communicative goal strength was higher in Show than in Do (Wilcoxon Signed-Ranks test: $Z = -5.04, p < .0001$). However, we did not find a difference in instrumental action informativeness (Wilcoxon Signed-Ranks test: $Z = -0.98, p = .33$). These analyses provide further confirmation that our model of belief-directed planning captures the qualitative dimensions of people's communicative demonstrations.

Experiment 2b: Learning Causal Structure from Demonstrations

Following the same structure as Experiment 1b, we tested whether Show demonstrations better conveyed information than Do demonstrations, as well as whether Communicative or Non-Communicative instructions influenced learning. We found that Show demonstrations were more effective at conveying the correct causal structure, but we found no effect of observer interpretation, consistent with the simulation results.

Materials, Procedure, Simulations. Three-hundred and twenty participants (150 female, 168 male, 2 neither) were recruited via Amazon Mechanical Turk to participate in our study. Two participants were assigned to each of the 80 demonstrators from Experiment 2a in two conditions (*Communicative* and *Non-Communicative*) using psiTurk (Gureckis et al., 2016). After completing a consent form, participants were shown instructions explaining that they would watch their partner play a game, that their goal was to reach the yellow square on each round and win 50 points, that red squares caused them to lose 10 points, and that green tiles were jumper tiles.

Participants observed their partner play 16 rounds of the game and had to determine whether the jumpers on that round were strong or weak. Each correct/incorrect answer was worth +/- 5¢. In only the Communicative condition they were told "Your partner knows that you are watching and is trying to show you whether the jumpers on that round are strong or weak."

Participants viewed a video of each demonstration as many times as they wanted (but at least once) and provided two judgments: whether the jumpers on that trial were strong or weak, and their confidence on a continuous slider ranging from "No Confidence" to "Extremely Confident". After completing all 16 trials, participants were asked several post-task questions. Procedures were approved by University of Wisconsin-Madison ED/SBS IRB (Study #2017-0830, title: "Studying human and machine interactions").

Using the same parameters as in the Experiment 2a simulations, we calculated both literal observer and pragmatic observer beliefs for Do or Show trajectories from Experiment 2a. Aggregated results are shown in Figure 9.

Results. We analyzed participants' strong/weak judgments and confidence ratings using mixed-effects models. We coded strong/weak judgments as correct or incorrect. A mixed-effects logistic regression was fit with correctness as the predictor variable; item (i.e. transition function and grid configuration), participant, and trial number intercepts as random effects; and demonstrator condition (Do/Show), ob-

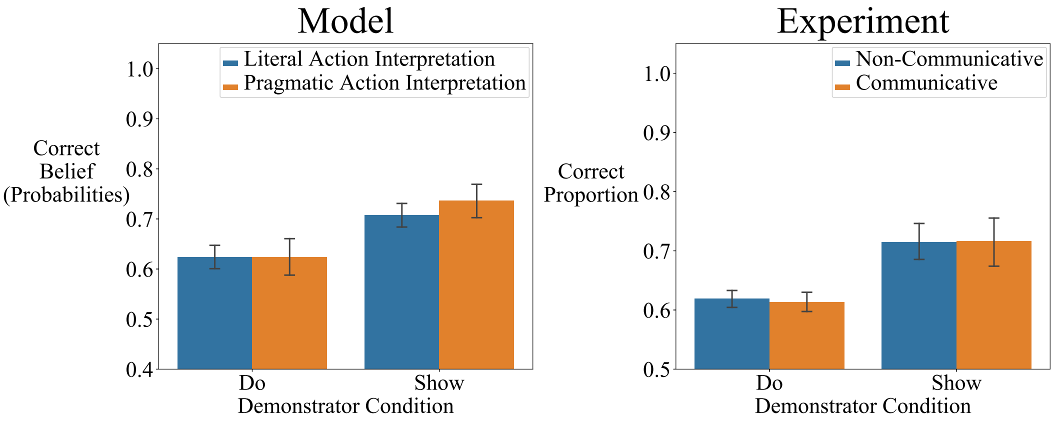

Figure 9. Experiment 2b model and human observers learning from human demonstrations. Observer models are either literal or pragmatic. Error bars are bootstrapped 95% confidence intervals.

server condition (Communicative/Non-Communicative), and their interaction as fixed effects. We found a significant effect of demonstrator condition corresponding to Show demonstrations increasing correctness by 2.31 times ($\beta = 0.90$, $SE = 0.18$, $Z = 5.11$, $p < .0001$ [Wald Z test]). However, there was no effect of observer interpretation ($\beta = -0.01$, $SE = 0.10$, $Z = -0.13$, $p = .90$ [Wald Z test]) and no interaction ($\beta = 0.08$, $SE = 0.21$, $Z = 0.38$, $p = .70$ [Wald Z test]).

For confidence judgments, we fit a mixed-effects linear model with confidence as the predictor variable; item, participant, and trial number intercepts as random effects; and observer condition, demonstrator condition, and their interaction as fixed effects. We found no effect of either demonstrator condition, observer condition, or their interaction.

Discussion

People can intentionally communicate causal knowledge through their actions. Experiment 2 tested whether people did this consistent with belief-directed planning using a Gridworld paradigm with strong or weak jumper tiles. We find that people are willing to use jumping and risky jumping as a signal for causal structure, and that belief-directed planning explains the behavior of Show participants more than Do participants. Risky jumping, in particular, provides strong evidence for belief-directed planning since it relies on the observer's beliefs about a demonstrator's aversion to risk and their causal knowledge. Finally, we find that observers more successfully learn causal structure from Show rather than Do participants, consistent with our model.

The current results further illustrate the generality of our framework for modeling communicative demonstrations. Similar to Experiment 1, we find that the model explains differences between doing and showing behaviors and their interpretation. We note that unlike in Experiment 1, we did not find that the framing of the trajectories as communicative or not influenced observer inferences. This may be due to the fact that the space of possible causal structures was smaller (two versus eight) and that the relative effectiveness of Do versus Show demonstrations left little room for observer interpretation to have an effect. In the next section, we analyze previous developmental studies in which infant observers were given experimentally controlled demonstrations in communicative or non-communicative contexts. There we find cases in which observer interpretation has a large influence on inferences.

Infant and Child Observer Studies

The previous adult experiments illustrate how belief-directed planning and pragmatic interpretation facilitate powerful forms of teaching and social learning. At the same time, the developmental literature documents a range of findings on the interpretation of communicative demonstration and their relation to learning action-guiding representations (Brugger et al., 2007; Southgate et al., 2009; Király et al., 2013; Hernik & Csibra, 2015; Buchsbaum et al., 2011; Butler et al., 2015; Sage & Baldwin, 2011; Hoehl et al., 2014). Having formalized the actor and observer roles in communicative demonstrations, we next compare the major qualitative predictions of our model against previous developmental findings.

Specifically, we revisit three studies in which an infant or child observed experimentally controlled demonstrations. Each set of studies focused on a different type of action-guiding representation that could be conveyed demonstratively: Király et al. (2013) focus on differential imitation of subgoals, Butler and Markman (2012) focus on learning generic causal properties, while Hernik and Csibra (2015) focus on inferring novel functional properties of tools. In each case, what is being conveyed is a decision-making representation that can be directly reflected in intentional action and thus demonstration.

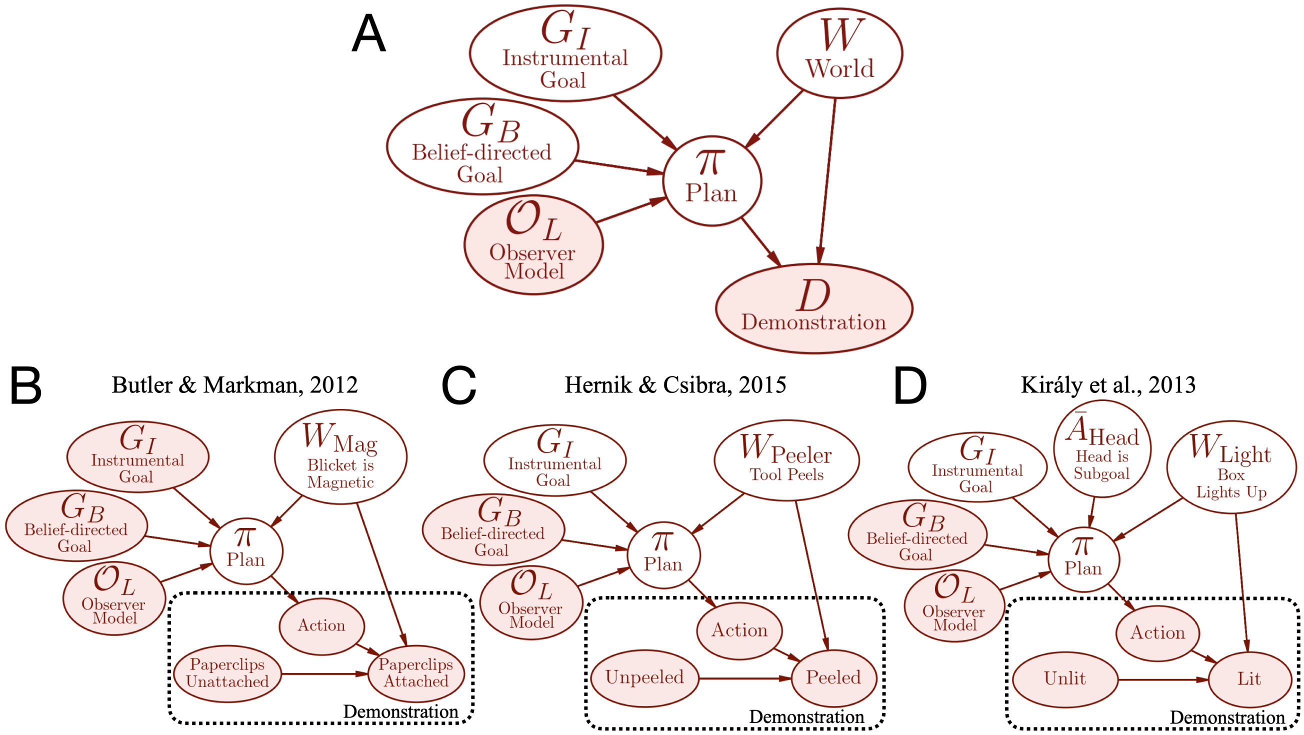

Figure 10. Comparison of general model with three developmental studies on interpreting communicative demonstrations. (A) Bayesian network diagram of pragmatic action interpretation (reproduced from Figure 1D) illustrating how inference from an observed demonstration is mediated by communicative planning. Shaded nodes indicate variables that are assumed or observed by the observer; open nodes indicate variables with uncertain values; edges indicate functional dependencies. (B, C, D) Diagrams illustrating how the variables in our general model map onto the specific experiments by Butler and Markman (2012), Király et al. (2013), and Hernik and Csibra (2015).

Additionally, the studies all compare learning between at least two observer conditions: A Communicative condition where demonstrations were performed following an ostensive cue, and an Intentional condition where there was no such cue. These conditions map directly onto our formalization of literal versus pragmatic action interpretation. By formulating participants as one or the other type of observer, we can can generate distinctive patterns of inferences and attributions. In general, we find that the pragmatic observer models tend to differ from literal observer models in two qualitative ways. First, we find *inferential amplification*, in which actions that are mildly diagnostic of instrumental goals become highly diagnostic of belief-directed goals. Second, we observe *deviation attribution*, in which actions that would have been attributed to irrationality or noise when interpreted as purely instrumental become attributable to belief-directed goals. These patterns are readily characterized in the framework of probabilistic inference and planning we have articulated, underscoring the value of our approach for understanding previous results.

Our primary goal is to test whether our account is compatible with the basic qualitative dimensions of existing findings and show how our model can be applied to interpreting studies of social learning of action-guiding representations. Previous work has used Bayesian models of cooperative communication to integrate results across studies and understand developmental changes in phenomena such as epistemic trust (Eaves & Shafto, 2017). Much of this work has focused on social learning in the context of labeling, and an important challenge is extending such analyses to settings that involve action-guiding representations (e.g., subgoals, object affordances, and tool functions). Although we do not address questions of developmental changes, the application of our model illustrates how constructs from the planning and reinforcement learning literature (Dayan & Niv, 2008; Sutton & Barto, 1998) can be used to define linking functions that connect model constructs to measures used to assess the learning of action-guiding representations. In the next few sections, we discuss how our model explains the qualitative findings of several studies. The full details of our formalization are included in the supplementary materials and implemented as probabilistic programs[1].

Throughout, we are required to make formally explicit several elements of the original experiments: the environmental context, the communicative context, and the demonstrations themselves. In deciding what aspects of a study are theoretically relevant, we have attempted to stay true to the original interpretations of the researchers. This illustrates how our approach can formally describe the key qualitative findings from the literature.

Learning Causal Structure from Demonstrations

Summary of Findings. Butler and Markman (2012) (Experiment 3) investigated how 3- and 4-year-olds learned about a novel causal property by observation, and how this depends on the communicative context. In their paradigm, participants observe an experimenter clean up a set of objects (these had been made messy in a prior distractor tasks, ostensibly the main task). The objects included many paperclips as well as a novel object (a wooden block with tape on it) that, earlier, had been labeled a "blicket". At the critical point of the experiment, the experimenter moves the blicket on top of the paperclips and they adhere to it by magnetic force.

The "demonstration" occurred in one of three experimental conditions: Accidental, in which the blicket was apparently dropped on the paperclips while being put away and the experimenter exclaimed "Oops!"; Intentional, in which the experimenter appeared to purposefully place the blicket on the paperclips without engaging the child; and Communicative, in which he addressed the child ("Look, watch this") before placing the blicket on the paperclips (Figure 11a). The children were then given a set of *inert* (i.e., non-magnetic) blickets and paperclips to play with.

Their main analyses revealed two important patterns of results for the 4-year-olds (but not 3-year-olds). First, those in the Communicative condition showed greater exploration and pickup-attempts than those in the Intentional and Accidental conditions. Second, there was no detectable difference in exploration or pickup-attempts between the Accidental and Intentional conditions (Figure 11D). These results indicate an influence of communicative context on how observers draw inferences.

Model Results and Discussion. The model captures the relationships between exploration/pickup-attempts in the Accidental, Intentional, and Communicative conditions in terms of literal and pragmatic interpretation. As illustrated in Figures 11A-C, we model observers as reasoning about whether or not blickets are magnetic (i.e., whether paperclips tend to attach to them) while also assuming that the observer initially understands the event as one in which the demonstrator has goal of putting the blicket away. We model observer inferences in the Accidental and Intentional conditions both as literal action interpretation, but model the Accidental demonstrator as "slipping" while attempting to put the blicket away. Inferences in the Communicative condition are then modeled as resulting from pragmatic interpretation. Figure 11C shows the results for a single parameterization of the model, but we found that the overall pattern of results was robust. Complete details on the formalization and alternative parameterizations can be found in the supplementary materials.

To understand the model, first consider the Intentional and Communicative conditions. Leading up to the critical part of the experiment, the participants can infer that the demonstrator has the goal of putting the blicket away since they had just put all the other objects away. Then at the critical part, the demonstrator places the blicket on the paperclips, causing them to stick together. Interpreted literally, the action appears as unexplained noise since it is irrelevant to putting the blicket away. However, the resulting observation that the paperclips stick to the blicket is informative: It provides evidence that blickets are magnetic. This is important because when the same demonstration is interpreted pragmatically, the act of placing the blicket on the paperclips and the observation that they stick together can be explained in terms of belief-directed intentions. In particular, the formerly unexplained action can be attributed to informative goals, while the possibility that the evidence for blicket-magnetism is intentional strengthens the inferences drawn from that very evidence. In other words, pragmatic interpretation leads to stronger inferences about blicket-magnetism due to deviation attribution and inferential amplification.

Our model can also explain why the Intentional and Accidental conditions did not differ. Specifically, in both conditions, the blicket landing on the paperclips is not relevant to the instrumental intention to clean the table—in both cases, this event is interpreted as noise. A difference is the source of the noise: In the Intentional condition, the noise is internal to the demonstrator's decision process, while in the Accidental condition the noise is "external"; for whatever reason, the demonstrator's hand slipped. Nonetheless, in both cases, the resulting observation itself provides only unintended evidence for blicket-magnetism.

Inferring Novel Tool Functions from Demonstration

Summary of Findings. Hernik and Csibra (2015) examined how infants could learn about novel tools and their functions. In a series of studies infants observed familiarization training videos in which a demonstrator manually used novel tools (e.g., a pink or blue flower pot turned upside down) on objects (e.g., a peeled or unpeeled banana). Their first study has two key features. First, these training demonstrations were marked as communicative by the demonstrator. Second, the objects were apparently transformed by a tool (e.g. an unpeeled banana, placed briefly under a blue tool, became peeled; a peeled banana, placed briefly under a pink tool, became unpeeled—healed). During the test trials they attempted to diagnose what the children had learned

[1] All models in this section were implemented using WebPPL (Goodman & Stuhlmüller, 2014) and can be found at `https://github.com/markkho/comdem-data-code`.

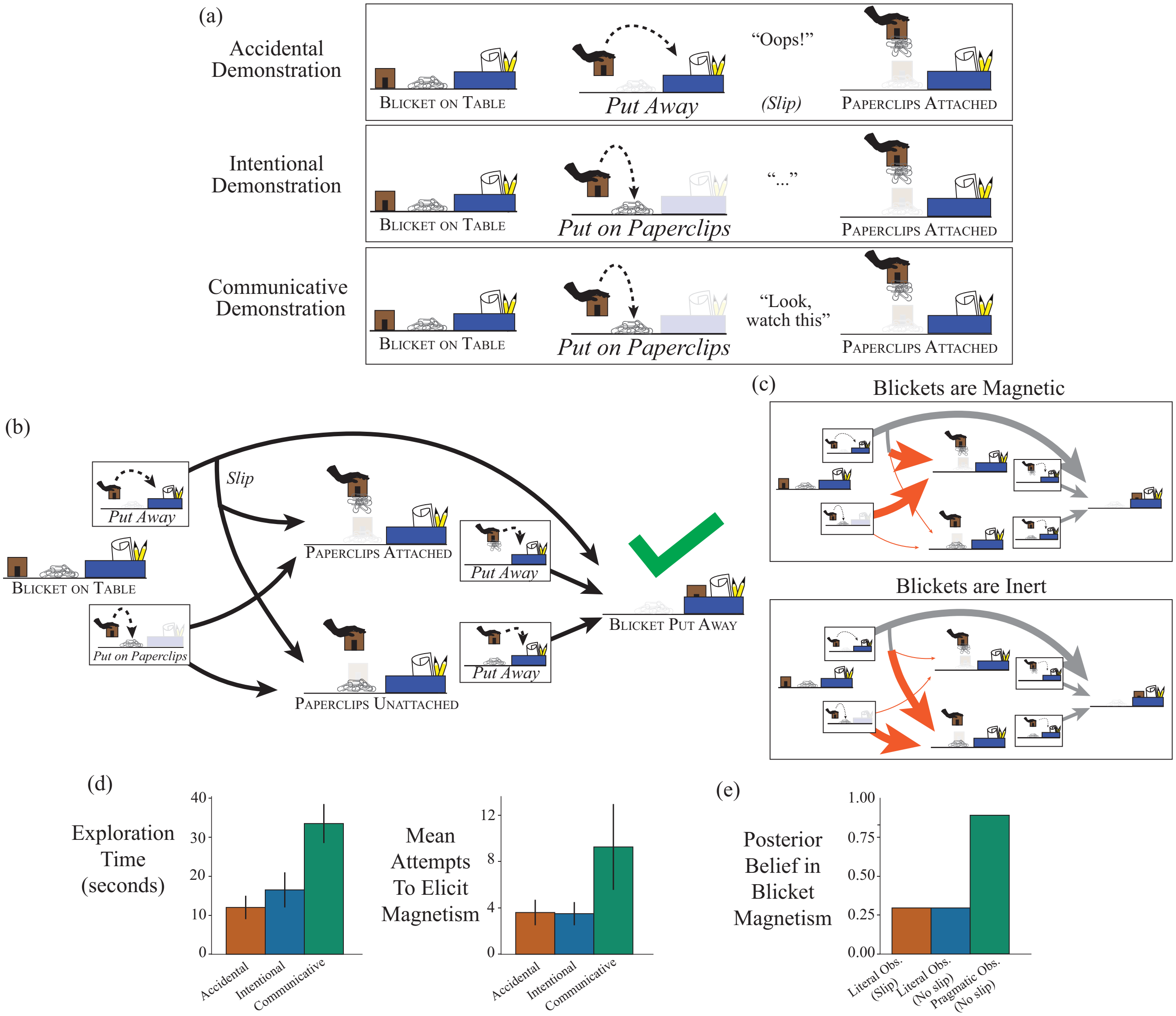

Figure 11. Summary and model of Butler and Markman (2012), Experiment 3. (a) Participants were shown a demonstration of the blicket (brown block) landing on the paperclips and sticking in one of three conditions: Accidental, Intentional, or Communicative. (b) Task model with possible transitions to and from states (Blicket on Table, Paperclips Attached, Paperclips Unattached, and Blicket Put Away) when taking actions (*Put Away* and *Put on Paperclips*). The only instrumental goal is putting the blicket away (green check). (c) Two possible causal relations correspond to blickets being magnetic or inert. Arrow width indicates relative probabilities. These differ based transition probabilities to Paperclips Attached and Paperclips Unattached (highlighted in orange). (d) Empirical results. Four-year-olds explored blickets more and attempted to elicit magnetism more in the Communicative condition. (e) The pragmatic action interpretation model (green) reasons about the demonstrator's informative goals, which leads to an amplification of the inferences produced by the literal observer model (blue).

about what the tools do. In order to do this, they showed infants videos of each tool while it was in use, such that the initial condition of the object was not observed. This was done without any communicative marking. On *congruent* trials, the final state of an object was congruent with that of the training trials for a tool (e.g., a peeled banana for a blue tool). On *incongruent* trials, the resulting state was the same as the initial state typically observed in the training phase for a tool (e.g., an unpeeled banana for a blue tool). Critically, they found significantly greater looking times on incongruent test trials, indicating that infants' expectations were violated when the result state of the tool differed from those of the training trials (Figure 12A, first row).

The authors also reported two additional comparison studies[2]. In the first of these, demonstrations were still communicative and child-directed, however, tools *did not transform* objects in the training trials. Unlike in the original study, they found no detectable difference between congruent and incongruent test trials. In the second comparison study, demonstrations were no longer child-directed and the tools transformed the objects in the training trials (as in the first study). Here, although they found a difference in congruent and incongruent looking times for initial test trials, this difference did not persist past the first set of test trials, unlike in the original study. Collectively, these three studies (summarized in Figure 12A) suggest that communicative marking and state changes interact when encoding novel tool functions, leading to especially robust, context-sensitive learning.

Model Results and Discussion. We model the observer in a single experimental trial as reasoning about whether the novel tool has the function of being a "banana peeler" or not (Figure 12B). Overall, we find a close correspondence between experimental looking times reported by Hernik and Csibra (2015) and the surprisal (negative log-likelihood) values for the literal/pragmatic observers with different training sequences (Figures 12C-D). In particular, the modeling accounts for two central features of the results: The difference in sustained violation of expectation between Studies 1 and 3, and the absence of a violation of expectation in Study 2. According to our account, these are explained by the amplification of inferences that result from pragmatic action interpretation.

First, consider the difference between Studies 1 and 3: The transformation demonstrations in Study 1 are performed in a child-directed communicative context, while those of Study 3 are not. (In other words, they correspond to the Communicative and Intentional conditions, respectively, of Király et al., 2013 and Butler & Markman, 2012). Thus, we model Study 3 participants as engaging in literal action interpretation when observing the transformation sequence. Although this allows them to draw a weak inference that the novel object is a banana peeler, it is easily defeated given contrary evidence, and so fails to drive robust violation of expectation in the test phase. In contrast, in Study 1, the communicative context makes the demonstrator's belief-directed intentions apparent, and so we model the participants as reasoning pragmatically. A pragmatic observer recognizes that evidence for the tool being a banana peeler has been intentionally presented to them, which leads to an amplification of the literal inference. In short, compared to participants in Study 3, those in Study 1 would reason that not only does the novel tool coincide with the change in the object, but the demonstrator wants the observer to know that this is a reliable feature of the world.

Meanwhile, our treatment of Study 2 uniquely draws out an important aspect of our model of pragmatic action interpretation: Although a communicative context is established, the demonstrator does not perform an action with any clear instrumental purpose. Put simply, the tool does not do anything (alternatively, one might say, the child fully expects that a random novel tool will not peel a banana, and thus no information is conveyed). In principle, the communicative context would allow for inferences to be amplified, but in this case it fails because there is nothing obvious to amplify.

Notably, our model actually identifies Study 2 congruent trials as more surprising than incongruent trials because the demonstrator's attempts to use the tool suggests that they expect it to change, even though it does not. (In other words they are surprised not by the state of the banana, but by the persistence of the demonstrator). Indeed, although Hernik and Csibra (2015) found no significant difference between congruent/incongruent looking times, they report that more than half of the infants tended to look at the congruent events more than the incongruent events.

Imitating Subgoals based on Communicative Demonstrations

Summary of Findings. Experiment 1 of Király et al. (2013) examined children's imitation of goal-directed behavior. In the modeling phase, infants observed an experimenter sit down and then bend over to use their head to touch a novel object, causing it to light up. This demonstration was performed in a 2 × 2 design. The first factor was whether the context was communicatively cued or not. In the Communicative conditions, the experimenter looked at the infant, called their name, and made sure they were paying attention before the demonstration. In the Intentional[3] conditions, the demonstrator did not interact with the infant, but waited until a signal was given from another experimenter that the infant was paying attention before performing the demonstration. The second factor manipulated whether the demonstrator's hands were occupied or free (Figure 13A). In the Hands-Occupied conditions, the demonstrator was wearing a blanket and clutching it with their hands. In the Hands-Free conditions, they were wearing a blanket but their hands were placed on the table next to the novel object.

During the test phase children had the opportunity to interact with the novel object. The main analysis examines whether the infants imitated the demonstrator by attempting to turn the light on with their head based on the two factors. Neither main effect was significant, but the interaction was significant. Specifically, in the Communicative condition, there was more imitation of the head action in the Hands-Free condition than the Hands-Occupied condition, whereas

[2] Hernik and Csibra (2015) report a fourth study that conceptually replicates the results of Studies 1 and 3. The analysis of Study 4 in terms of our framework is identical to that of Studies 1 and 3.

[3] Király et al. (2013) use the term "Incidental" to describe this condition.

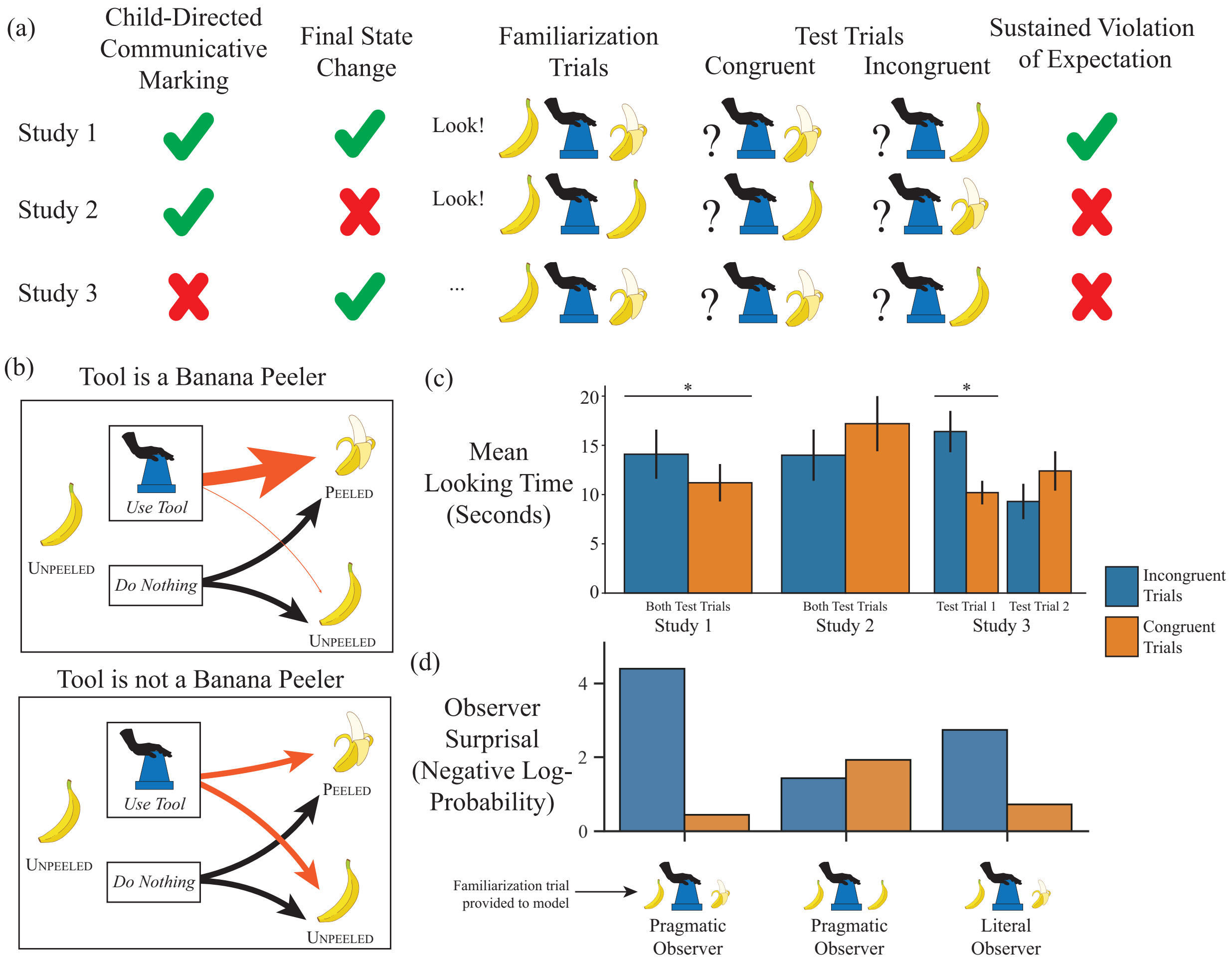

Figure 12. Summary and model of Hernik and Csibra (2015) Experiments 1, 2, and 3. (a) Participants viewed video trials where novel tools (blue or pink flower pots) were used on bananas (peeled or unpeeled). The banana either changed or not when the tool was used. Note that the event sequence for a single tool here. Violation of expectation (VOE) measures were used to assess whether a novel functional concept (e.g. the tool is a banana peeler) was learned. VOE was sustained across multiple test trials only when training occurred with communicative marking and target transformation. (b) Possible causal structures in our model of a single trial. It is possible that the banana changes from a state Unpeeled, to a new state, Peeled, regardless of what action is taken for some (unspecified) reason unrelated to the flower pot. If the tool is a banana peeler (top), then using it makes a transformation more likely. If it is not (bottom), then tool use has no additional effect. Note that although the participants never see the banana change without the tool, the model considers the possibility that the tool has no effect. (c) Looking-time results from studies 1-3. (d) Model posterior surprisals for congruent/incongruent test trial observations. A pragmatic observer trained on a tool-transformation sequence (left) has a high surprisal on an incongruent trial. The same observer model trained on a non-transformation sequence (middle) expects both sequences nearly equally. A literal observer trained on a tool-transformation sequence (right) has a higher surprisal on the incongruent observation, but lower than the first pragmatic observer.

there was no detected difference in the Intentional condition (Figure 13D).

Intuitively, a person who turns on a light with their head in the Hands-Occupied condition uses their head only because their hands are occupied, whereas a person who turns on a light with their head in the Hands-Free condition uses their head because it is necessary or preferable. Moreover, a person who communicatively demonstrates turning on a light with their head in the Hands-Free condition is choosing a highly diagnostic signal that head-use is important or preferable. Our model naturally captures these intuitive principles.

Model Results and Discussion. The modeling captures two key patterns in the results: (1) In Hands-Free, head-use imitation increases from the Intentional to Communicative conditions, and (2) in Hands-Occupied, head-use imitation decreases from the Intentional to Communicative conditions. Our model explains how these effects arise naturally when an observer engages in pragmatic action interpretation in the Communicative conditions.

Figures 13B-C illustrate our formalization of the task. A key aspect of the experiment we model is the difference

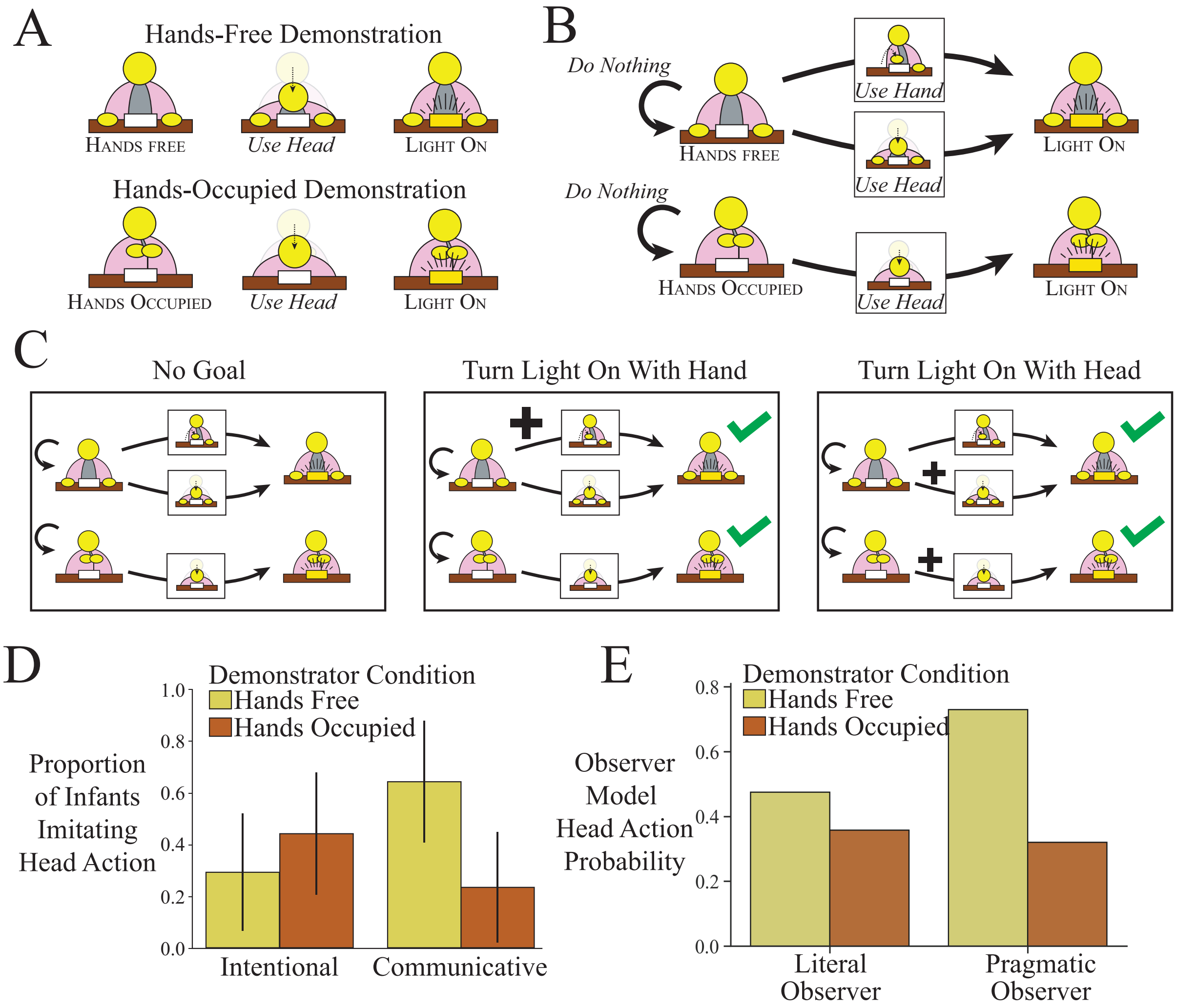

Figure 13. Summary and model of Király et al. (2013), Experiment 1 results. (a) Participants observed an experimenter use their head to light up a novel object in one of two conditions: Wearing a blanket (Hands-Free) versus holding a blanket (Hands-Occupied)—and in a Communicative or Intentional condition (not shown). (b) A minimal model of the environmental constraints - The demonstrator can start in the Hands Free state and deterministically transition to the Light On state by taking either *Use Head* or *Use Hand*, or *Do Nothing*. Alternatively, they can start in Hands Occupied and take *Use Head* or *Do Nothing*. (c) Space of possible utility functions in model, corresponding to subgoal/goals of the demonstrator. Plus signs indicate actions with a stronger bias (e.g., lower cost for achieving the goal). Checks indicate a +1 reward. (d) Empirical results in Király et al. (2013). Infants differentially attempt to use their head in the Communicative conditions but not the Intentional conditions. Error bars are 95% binomial confidence intervals. (e) Model results. After observing either the Hands-Free or Hands-Occupied demonstration, each observer model has a belief over the three possible utility functions, which induces an expected reward function. This plots the observer's softmax probability ($\frac{1}{\alpha} = 2.5$) of taking *Use Head* from Hands Free under the expected rewards. In particular, our models capture the exaggerated difference in the communicative conditions.

between the available actions in Hands-Free versus Hands-Occupied. In Hands-Free, both *Use Hand* and *Use Head* are available, whereas in Hands-Occupied, only *Use Head* is available. The differential structure of available actions directly affects what evidence is provided about what actions are optimal. This evidence, in turn, interacts with whether the observer interprets actions literally or pragmatically. We discuss this process in the Hands-Free conditions and then the Hands-Occupied conditions.

In the Hands-Free conditions, the demonstrator can use either his hands or his head. When a literal observer sees the demonstrator use his head to turn on the light, she will assume that this is either because using one's head is optimal, or because the demonstrator acted suboptimally (which can always occur with some low probability). This only causes a slight increase in her belief that head-use is optimal since her prior belief was low and there remains the possibility that the demonstrator acted suboptimally. Notably, however, a pragmatic observer will draw stronger inferences. This is because she assumes that the demonstrator used his head not only for instrumental reasons (e.g., to efficiently turn on the light), but also belief-directed reasons (e.g., to provide diagnostic information). Put informally, the communicative context leads to amplification of existing inferences: The pragmatic observer knows that the demonstrator wants the literal observer to have certain inferences, which strengthens those inferences. (It is as if the pragmatic observer reasons: "He wants me to think that head-use is optimal, so head-use really must be optimal"). The communicative context also offers an explanation for seemingly suboptimal behavior: The literal observer only attributes head-use to suboptimality or the low-probability case of head-use being optimal. In contrast, the pragmatic observer can attribute head-use to the belief-directed goal to inform her that head-use is optimal.

In contrast to the Hands-Free conditions, in the Hands-Occupied conditions, the demonstrator can only use his head. (This is known to the observer, who can see that the demonstrator's hands are occupied). To a literal observer, head-use provides evidence that turning the light on is a goal, but it does not imply that head-use is optimal in general. Rather, it provides evidence that one should turn on the light, even if it requires using one's head. Interpreted pragmatically, the overt evidence for this becomes stronger evidence that turning on the light is a goal (i.e., inferential amplification), leading to the strong inference that one should turn on the light (which is typically done with one's hands).

Given the resulting observer inferences about goals and subgoals, we simulate imitation by calculating how they would act in Hands-Free. Figure 13E plots the probability of the literal and pragmatic observers taking *Use Head* after observing the demonstrator actions in Hands-Free versus Hands-Occupied for a particular set of parameters. In the supplementary materials, we describe the formalization of the task in more detail and provide an analysis across a range of parameterizations. Overall, we find that the pragmatic action interpretation model is able to capture the qualitative pattern of results reported in the original experiments.

General Discussion

How do people communicate with their actions? To address this question we combine ideas from work on inverse planning and linguistic pragmatics. Specifically, we develop and test an account of communicating with one's actions in terms of belief-directed planning and pragmatic action interpretation. Our account formalizes two ideas: First, communicative demonstrations are grounded in shared assumptions about the interpretation of instrumental action. That is, both demonstrator and observer understand instrumental planning and inverse planning. Second, communicative demonstrators rationally plan their actions based on their model of the environment and an observer's inverse planning. Pragmatic action interpretation then involves reasoning about the instrumental and belief-directed intentions underlying such demonstrations. We have shown how this model facilitates powerful forms of teaching and observational learning in theory and captures data from novel and existing experiments in practice.

In Experiments 1 and 2, we used our models to predict how people teach about novel tasks by engaging in belief-directed planning as well as how they learn via pragmatic action interpretation. Using a combination of simulations and model-fitting, we show a close correspondence between our account and human behavior and judgments. Additionally, we examined three previously reported developmental studies of learning from communicative demonstrations in order to assess the theoretical import of our models. Each set of studies focus on learning a different decision-making representation: Király et al. (2013) examine imitation of subgoals; Butler and Markman (2012) test learning about generic causal properties; and Hernik and Csibra (2015) study inferring novel tool functions. We show how these previously reported findings can be understood in terms of belief-directed planning and pragmatic action interpretation given particular contexts and environmental constraints.

In short, we have developed and tested a model for characterizing communicative demonstrations that combines ideas from work on language pragmatics and pedagogy (Frank & Goodman, 2012; Shafto et al., 2014) with work on value-guided decision-making and planning (Dayan & Niv, 2008; Newell & Simon, 1972). This adds to a growing body of computational work that characterizes the communicative aspects of non-verbal social interactions (Ho, Cushman, Littman, & Austerweil, 2019; Ho, MacGlashan, Littman, & Cushman, 2017). Additionally, our formal approach provides a complementary perspective to existing accounts of the evolution and development of cognitive abilities supporting human social learning (Tomasello et al., 2005; Csibra & Gergely, 2009). In the remainder of this section, we discuss the implications of our work for formal models of social cognition as well as our understanding of the cognitive mechanisms underlying human sociality.

Summary of Model Contributions

We have developed a framework for characterizing communicative demonstrations that combines ideas from instrumental planning with those from pragmatic reasoning. Our approach extends existing accounts in several ways by providing an account of how the meaning of communicative actions are grounded as well as how agents can reason about both the instrumental and belief-directed effects of actions.

Grounding Communication in Instrumental Action. Our model reveals connections between communicative demonstrations and other forms of communication, such as language and teaching by example. For instance, in Rational Speech Act (Frank & Goodman, 2012; Goodman & Frank, 2016; Yoon, Tessler, Goodman, & Frank, 2017) and Bayesian Pedagogy (Shafto et al., 2014) models, a transmitter (e.g. a speaker or teacher) provides a signal (e.g. an utterance or an example) to a receiver (e.g. a listener or learner) who must infer an underlying message (e.g. a linguistic meaning or novel concept). In these settings, candidate signals have a default interpretation, which communicative partners can use to recursively anticipate each others' selection of a particular signal and interpretation of that signal. In models of linguistic pragmatics, this default interpretation is the semantics of words, while in models of concept teaching, it is a probabilistic concept class. One way to identify such interpretations is to estimate them empirically, for instance, by asking people what the expected semantics are in non-pragmatic contexts (Frank & Goodman, 2012; Kao, Wu, Bergen, & Goodman, 2014). Alternatively, one can derive constraints on default interpretations assuming the cooperative interpretation is optimal (Yang et al., 2017). From a theoretical perspective, the default interpretation of signals plays a critical role in determining how interactants can coordinate on the meaning of a communicative act.

Our model illustrates a new and powerful form of default interpretation: instrumental action and inverse planning. In our model, the default interpretations are determined by relating possible environments that an agent occupies to a theory of instrumental action within that environment. That is, it is derived from value-guided decision-making, a general framework for describing the behavior of any adaptive system or organism (Sutton & Barto, 1998; Newell, 1982; Anderson, 1990). Prior work has established that the capacity to recognize intentional behavior is present in humans from a young age (Gergely & Csibra, 2003; Malle, 2008). In recent years, various aspects of mindreading, including reasoning about beliefs, desires, intentions, uncertainty, appraisal, and emotions, have been cast as probabilistic inference (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Ong, Zaki, & Goodman, 2015). Put simply, we propose that people can ground inferences about what a person is trying to communicate in inferences about what she is trying to do.

Reasoning about the instrumental and belief-directed effects of actions. Instrumental actions are not inherently communicative, but people can nonetheless use them communicatively. For example, consider a cyclist who takes her hands off of the handlebars as she is riding. In itself, this is a meaningless physical action. But, in the right context, it acquires communicative meaning: Suppose that the cyclist is riding beside her friend in the park and, while her friend is watching, takes her hands off the handlebars without falling over. In doing so, she is effectively saying, "I don't need my hands to stay balanced!"

Our model explains what kind of context allows inherently "meaningless" actions to become "meaningful" via a process of recursive planning and inference. Specifically, demonstrators plan over the instrumental and belief-directed effects of their actions, while observers infer the instrumental and/or belief-directed intentions behind a demonstrator's actions. We characterize these processes by merging a formalization of sequential decision-making (Newell & Simon, 1972; Puterman, 1994) with a formalization of recursive theory of mind (Camerer et al., 2004; Baker et al., 2009).

By focusing on the interplay of instrumental and belief-directed planning, this work builds on several existing models and suggests new directions for research. In particular, by introducing inverse planning as a component of a learner's inference model, we extend work by Shafto et al., 2014 (Experiment 3) that modeled how people teach causal concepts, as well as work by Buchsbaum et al., 2011 that modeled teaching action sequences. Related work by Rafferty et al., 2016 models the role of planning over learner beliefs to teach concepts and considers how different models of a learner's belief dynamics affect teaching strategies. Here, we briefly explored observer uncertainty about the demonstrator (e.g., the uncertain pragmatic observer who reasons about whether demonstrators are communicative), but characterizing demonstrator uncertainty about the observer in instrumental planning settings is a promising direction for future work.

Connections and Future Directions

Communicative demonstrations play a key role in teaching and social learning as well as a range of human social interactions. Moreover, as we have argued, the mechanisms underlying communicative demonstration can be understood in terms of familiar ideas from probabilistic inference (Tenenbaum & Griffiths, 2001), model-based planning (Dayan & Niv, 2008), and cooperative communication (Shafto et al., 2014; Goodman & Frank, 2016). Here, we discuss how our account and findings connect to other active areas of research.

First, while there is extensive research into the mechanisms of standard instrumental planning, belief-directed planning has been less systematically explored. In our discussion of the model, we noted how recursive reasoning and planning are both computationally demanding processes. A promising approach will be to ask whether, and how, the key features that enable efficient and powerful instrumental planning in physical domains might also extend to the case of belief-directed planning. These include hierarchical action representations (Botvinick, Niv, & Barto, 2009; Barto & Mahadevan, 2003), state abstractions (Ho, Abel, Griffiths, & Littman, 2019), and function approximation (Sutton, McAllester, Singh, & Mansour, 2000). Additionally, by better understanding the computational processes involved in belief-directed planning, we can ask whether it relies on the same cognitive and neural substrates as instrumental planning, or instead on analogous but distinct mechanisms and representations.

Second, our planning and inference models underscore the flexibility of human social learning mechanisms, which can help us understand the distinctive scale and scope of human sociality and culture. For example, there is an emerging consensus that both human children and certain non-human primates can imitate in the strong sense of copying intentions (Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009). Nonetheless, only humans appear to engage in "overimitation" whereby seemingly causally irrelevant actions are still copied by an observer with high fidelity (Lyons, Damrosch, Lin, Macris, & Keil, 2011). Thus, precisely characterizing the mechanisms underlying overimitation will be key for understanding human cultural transmission (Keupp, Behne, & Rakoczy, 2018; Clay & Tennie, 2018). Our account extends previous computational proposals that attempt to provide a rational interpretation of overimitation (e.g., Buchsbaum et al., 2011). Specifically, because our framework allows for reasoning about whether an actor has communicative intentions, what their content might be, and how they plan to convey this information, it can capture different pragmatic inferences than those of previous accounts. For instance, interpreting an action as an ostensive cue requires recognizing that the demonstrator is first establishing they have an intention to communicate something and then taking actions to communicate it. Put another way, our framework could be used to derive ostensive cues as components of a larger plan whose ultimate goal is to convey information.

Additionally, our account naturally raises the question, "What kinds of belief-directed goals do people tend to have", and the complementary question, "What kinds of belief-directed goals do we tend to expect of others?" Put differently, what are the actual and subjective priors on communicative intentions? Some existing theories take a strong stand on these questions. For instance, the theory of natural pedagogy (Csibra & Gergely, 2009) proposes that child observers treat communicative demonstrations as conveying *relevant, generalizable information* (e.g. "blickets are magnetic"), and a number of findings support this view (Butler & Tomasello, 2016; Butler & Markman, 2012). Our approach suggests a way to understand why this might be the case: Relevant, generalizable knowledge is useful, adults have lots of it, and infants and children have much less. If an adult initiates communication with a child, the theory of natural pedagogy embodies a very natural prior on the adult's communicative intentions. But different settings and social relationships might imply very different distributions over communicative intentions, as we consider next.

We have focused on communicative demonstrations aimed at teaching skills, but this model applies to a much larger array of human behaviors. Often, for instance, we use communicative demonstrations in arbitrary contexts to convey our feelings (e.g., giving roses on Valentine's Day), intellect (e.g., asking a very technical question during a department colloquium) or income (e.g., driving a Maserati). Such demonstrations are a form of costly signaling (Gintis, Smith, & Bowles, 2001) and have been studied in situations ranging from information search (Hoffman, Yoeli, & Nowak, 2015) to time-consuming deliberation (Jordan, Hoffman, Nowak, & Rand, 2016; Levine, Barasch, Rand, Berman, & Small, 2018) to third-party punishment (Millet & Dewitte, 2007; Fehrler & Przepiorka, 2013; Jordan, Hoffman, Bloom, & Rand, 2016). We would expect that many of the benefits of communicative demonstrations for teaching skills carry over into these other settings. For instance, the capacity to adaptively generate and interpret costly signals in novel contexts may play a key role in supporting complex forms of cooperation, coordination, and politics, much like how flexible teaching supports enhanced cultural accumulation (Tennie, Call, & Tomasello, 2009). Future research should explore the connections between the mechanisms we explore here, signaling behaviors in other domains, and the distinctive scale and scope of human sociality.

Conclusion

We have formulated and tested a computational account of communicative demonstrations based on rational, belief-directed planning and pragmatic action interpretation. The models we develop build on existing theoretical work and are supported by the results of novel experiments and previously reported findings. This account provides insight into the mechanism of human communication, imitation, and interaction while also suggesting future directions for examining the relationship between communicative demonstrations and other dimensions of human sociality.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

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), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, *13*(1-2), 41–77.

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–280.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental science*, *5*(1), 72–83.

Brugger, A., Lariviere, L. A., Mumme, D. L., & Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. *Child development*, *78*(3), 806–824.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child development*, *83*(4), 1416–1428.

Butler, L. P., Schmidt, M. F., Bürgel, J., & Tomasello, M. (2015). Young children use pedagogical cues to modulate the strength of normative inferences. *British Journal of Developmental Psychology*, *33*(4), 476–488.

Butler, L. P., & Tomasello, M. (2016). Two- and 3-year-olds integrate linguistic and pedagogical cues in guiding inductive generalization and exploration. *Journal of Experimental Child Psychology*, *145*, 64 - 78. doi: https://doi.org/10.1016/j.jecp.2015.12.001

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Cartmill, E. A., Beilock, S., & Goldin-Meadow, S. (2012). A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 129–143.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse studies*, *7*(4-5), 507–525.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324–347. doi: 10.1037/rev0000026

Clay, Z., & Tennie, C. (2018). Is overimitation a uniquely human phenomenon? Insights from human children as compared to bonobos. *Child development*, *89*(5), 1535–1544.

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, 201720963. doi: 10.1073/pnas.1720963115

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. doi: 10.1016/j.conb.2008.08.003

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

Eaves, B. S., & Shafto, P. (2017). Parameterizing developmental changes in epistemic trust. *Psychonomic bulletin & review*, *24*(2), 277–306.

Eaves Jr, B. S., & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. In *Advances in child development and behavior* (Vol. 43, pp. 295–319). Elsevier.

Fehrler, S., & Przepiorka, W. (2013). Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts. *Evolution and Human Behavior*, *34*(2), 139–145.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. doi: 10.1126/science.1218633

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. Institute for Logic, Language and Computation.

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. doi: 10.1016/0010-0277(95)00661-H

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of theoretical biology*, *213*(1), 103–119.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2018-9-12)

Grice, H. P. (1957). Meaning. *The philosophical review*, 377–388.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hernik, M., & Csibra, G. (2015). Infants learn enduring functions of novel tools from action demonstrations. *Journal of Experimental Child Psychology*, *130*, 176–192.

Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., & Epstein, R. (2005). Bootstrap Methods and Permutation Tests. In D. Moore & G. P. McCabe (Eds.), *Introduction to the Practice of Statistics.* New York: Freeman.

Ho, M. K., Abel, D., Griffiths, T. L., & Littman, M. L. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111 - 116. doi: https://doi.org/10.1016/j.cobeha.2019.05.001

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019, mar). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520–549.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 3027–3035). Curran Associates, Inc.

Ho, M. K., Littman, M. L., Cushman, F., & Austerweil, J. L. (2018). Effectively learning from pedagogical demonstrations. In C. Kalish, M. Rau, T. Rogers, & J. Zhu (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. XX–XX). Austin, TX: Cognitive Science Society.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*. doi: 10.1016/j.cognition.2017.03.006

Hoehl, S., Zettersten, M., Schleihauf, H., Grätz, S., & Pauen, S. (2014). The role of social interaction and pedagogical cues for eliciting and reducing overimitation in preschoolers. *Journal of Experimental Child Psychology*, *122*, 122–133.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q based and r-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (p. 11-42). Georgetown University Press.

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Comput Biol*, *11*(6), e1004254.

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. doi: 10.1016/j.cognition.2015.03.006

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99–134.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Keupp, S., Behne, T., & Rakoczy, H. (2018). The rationality of (over)imitation. *Perspectives on Psychological Science*, *13*(6), 678-687. doi: 10.1177/1745691618794921

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. L. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, *16*(2), 209–226.

Király, I., Csibra, G., & Gergely, G. (2013). Beyond rational imitation: Learning arbitrary means actions from communicative demonstrations. *Journal of experimental child psychology*, *116*(2), 471–486.

Levine, E. E., Barasch, A., Rand, D., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experimental Psychology: General*, *147*(5), 702.

Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, *66*(2), 81.

Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M., & Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*(1567), 1158–1167. doi: 10.1098/rstb.2010.0335

Malle, B. F. (2008). The fundamental tools, and possibly universals, of human social cognition. *Handbook of motivation and cognition across cultures*, 267–296.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380.

Millet, K., & Dewitte, S. (2007). Altruistic behavior as a costly signal of general intelligence. *Journal of research in Personality*, *41*(2), 316–326.

Munos, R., & Moore, A. (2002). Variable Resolution Discretization in Optimal Control. *Machine Learning*, *49*(2-3), 291–323. doi: 10.1023/A:1017992615625

Nassar, M. R., & Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54. doi: 10.1016/j.cobeha.2016.04.003

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104) (No. 9). Prentice-Hall Englewood Cliffs, NJ.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., & Castelfranchi, C. (2019). The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of life reviews*, *28*, 1–21.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (1st ed.). New York, NY, USA: John Wiley & Sons, Inc.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster Teaching via POMDP Planning. *Cognitive Science*, *40*(6), 1290–1332. doi: 10.1111/cogs.12290

Sage, K. D., & Baldwin, D. (2011). Disentangling the social and the pedagogical in infants' learning about tool-use. *Social Development*, *20*(4), 825–844.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*(2), 226 - 233. doi: https://doi.org/10.1016/j.cognition.2009.08.009

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436–447.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. doi: 10.1016/j.cogpsych.2013.12.004

Southgate, V., Chevallier, C., & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental science*, *12*(6), 1013–1019.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592.

Spector, B. (2007). Scalar implicatures: Exhaustivity and gricean reasoning. In A. B. M. Aloni & P. Dekker (Eds.), *Questions in dynamic semantics* (p. 225 - 249). Elsevier.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA, USA: Harvard University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems* (pp. 1057–1063).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1528), 2405–2415.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–690.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2417–2428. doi: 10.1098/rstb.2009.0069

Wingate, D., Goodman, N. D., Roy, D. M., Kaelbling, L. P., & Tenenbaum, J. B. (2011). Bayesian policy search with policy priors. In *Twenty-second international joint conference on artificial intelligence.*

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1431), 593–602.

Yang, S. C.-H., Yu, Y., Givchi, A., Wang, P., Vong, W. K., & Shafto, P. (2017). Optimal Cooperative Inference. *arXiv preprint arXiv:1705.08971*.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). “I won’t lie, it wasn’t amazing”: Modeling polite indirect speech. In *Proceedings of the thirty-ninth annual conference of the cognitive science society.*

Supplementary Materials

Details of Formalism

In this section, we describe the details of our modeling framework and implementation. The presentation is intended to be self-contained but builds on formal models of sequential decision-making and reinforcement learning in Markov Decision Processes (MDPs), reviewed in Puterman, 1994; Sutton & Barto, 1998. Code for the models and analyses are available at `https://github.com/markkho/comdem-data-code`.

Instrumental Planning and Acting

People can take intentional actions in order to achieve their goals. For instance, when riding a bicycle to work, one has the goal of reaching a destination while also minimizing the amount of pedaling one has to do. This requires having a model of the world (e.g., of how pedaling affects the wheels and which streets lead to work) as well as a goals (e.g., being at work, pedaling as little as possible). Planning involves using a world model to reason about what actions best realize one's goals and then enacting a plan.

Formally, planning and intentional action relies on a *world model* that captures causal knowledge about the world and *utilities* for different states of affairs. For a particular possible world $w \in \mathcal{W}$, a transition model $P(s' \mid s, a; w)$ is defined over an object-level state space $\mathcal{S}$ and describes how the environment probabilistically updates to a new state s' given a previous state s and an action a. An agent's instrumental goal maps states to utilities, $G_I : \mathcal{S} \to \mathbb{R}$. An agent's instrumental goals can have multiple components such as the goal to minimize action costs or use subgoals (e.g., reaching work while pedaling as little as possible).

Planning involves computing how well a sequence of actions realizes goals, given a model of the world. This can be represented by the *value* of an action, which is how much future expected utility one gains from an action, given that afterwards, one takes all the best actions. In general, this quantity is difficult to compute (Bellman, 1957), but we assume that in these relatively simple settings people can compute this quantity near optimally.

Formally, the *Q-value* of an action a taken from a state s in world w given instrumental goal G_I is represented by the following recursive equations:

$$
\begin{aligned}
&Q(a, s; w) = \\
&\quad \sum_{s'} P(s' \mid s, a; w) \left[G_I(s, a, s') + \gamma \max_{a'} Q(a', s'; w) \right],
\end{aligned}
\tag{5}
$$

where $\gamma \in [0, 1]$ is a discount rate that controls the relative weighting of temporally close and distant utilities.

Q-values express the goodness of actions, but a linking function is required to translate them into action probabilities. To allow for systematic deviations from perfect optimality, we use an *ε-softmax* decision rule that has been successfuly applied to modeling human decision-making in psychology and reinforcement learning (Luce, 1959; Nassar & Frank, 2016; Collins & Frank, 2018). The *ε-softmax* decision-rule has two parameters: a random choice probability ε and a softmax inverse temperature parameter α. Intuitively, the decision rule expresses randomly selecting any available action with probability ε or choosing an action that soft-maximizes the Q-value with inverse temperature parameter α. The action probabilities associated with a plan π are then:

$$
\pi(a \mid s; w) = (1 - \varepsilon) \frac{e^{\alpha Q(s,a;w)}}{\mathcal{Z}(s; w)} + \frac{\varepsilon}{|\mathcal{A}(s)|}, \tag{6}
$$

where $\mathcal{Z}(s; w) = \sum_a e^{\alpha Q(s,a;w)}$ is a normalizing constant and $|\mathcal{A}(s)|$ is the number of actions available at a state s.

Enacting a plan involves both the agent's plan and the actual dynamics of the world. In this work, we focus on how enacted plans lead to demonstrations that both the actor and observer are aware of. Formally, a demonstration is a sequence of states and actions, $D = (s_0, a_0, s_1, ..., s_{T-1}, a_{T-1}, s_T)$ that results from executing a plan π in the world w. The probability of a demonstration starting from a state s_0 is then:

$$
P(D \mid \pi, w) = \prod_{t=0}^{T} \pi(a_t \mid s_t; w) P(s_{t+1} \mid s_t, a_t; w) \tag{7}
$$

Inverse planning and literal observer models

We are interested in how observers interpret demonstrations, and what consequences this has for communication. The interpretation of intentional action has been successfully modeled as *inverse planning* (Baker et al., 2009), in which a generative model of planning is "inverted" to allow for inferences about what intentions gave rise to an observed sequence of actions. In our case, we are interested in how observers can draw inferences about the world by assuming actions are generated by a plan. Formally, this corresponds to doing Bayesian inference over the demonstration model expressed in Equation 7:

$$
\begin{aligned}
P(w \mid D, G_I) &\propto P(D \mid w, G_I) P(w) \\
&= \sum_{\pi} P(D \mid \pi, w) P(\pi \mid w, G_I) P(w)
\end{aligned}
\tag{8}
$$

As we discuss in the main text, this process of inverse planning can be used to define a *literal observer model O_L* by associating beliefs b with probability distributions that are updated according to Equation 8. Specifically, the one-step literal observer belief state updates upon observing a state, action, and next-state are given by:

$$O_L(b' \mid s,b,a,s') = \begin{cases} 1 & \text{if for all } w, \\ & b'(w) \propto \pi(a \mid s; w)P(s' \mid s,a;w)b(w) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We note that this formulation of belief-state transitions is analogous to techniques for transforming partially observable Markov decision processes (POMDPs) into fully observable belief-state Markov decision processes (Kaelbling, Littman, & Cassandra, 1998). The key difference is that we consider belief dynamics in another agent rather than in one's own belief space. Additionally, here we assume that observer belief dynamics are deterministic and known, but it would be straightforward to extend these ideas to richer observer inference models (e.g., see work by Rafferty et al., 2016).

Planning and Acting in Belief Space

Instrumental plans determine the optimal actions given a world model w and instrumental goals G_I. We can extend this logic to planning and acting in belief space by having Q-values additionally incorporate observer belief dynamics, O_L, and belief-directed goals, G_B. Formally, the Q-values for a belief-directed agent are:

$$\begin{aligned} Q(a,s,b;w) = & \\ \sum_{s',b'} P(s' \mid s,a;w)O_L(b' \mid s,b,a)\Big[& G_I(s,a,s';w) \\ & + \beta G_B(b',b;w) \\ & + \gamma \max_{a'} Q(a',s',b';w)\Big] \end{aligned} \tag{10}$$

where $\beta \in \mathbb{R}^+$ is a belief-directed goal weighting parameter. Note that when $\beta = 0$, the belief-directed Q-values are equal to the instrumental Q-values.

This formulation is general enough to express arbitrary belief-directed goals (e.g., wanting to hide one's intentions rather than show them). Here, we focus on belief-directed goals that involve increasing an observer's belief in the true state of the world:

$$G_B(b',b;w) = b'(w) - b(w). \tag{11}$$

Given the Q-values over joint ground and belief states (Equation 10), we can use the ε-softmax decision rule to determine the *belief-directed plan*, $\pi(a \mid s,b;w)$. Note that belief-directed plans, unlike instrumental plans, are determined by both the current state of the world s, as well as the observer's current beliefs, b.

Approximating Belief-directed planning. Here we describe the algorithmic details of our approximation procedure for solving a belief-directed plan for the Gridworld tasks (Experiments 1 and 2). To model planning in belief space, it is necessary to approximate the value function. We did this by constructing a discretized, point-based MDP with an approximate ground and belief-state transition function $\hat{P}(s',b' \mid s,b,a)$ (Munos & Moore, 2002). We discretized the original belief-state space to a set B_D and constructed a transition function where, for each $a \in A$, $s \in S$, and $b_D \in B_D$, $\hat{P}(s',b'_D \mid a,s,b_D) = \sum_{b'} P(s' \mid s,a;w)O_L(b' \mid b_D,s,a,s')NN(b',b'_D)$, where $NN(s',s'_D)$ is an indicator function for whether out of the points in B_D, b'_D is the nearest neighbor of b'. This then serves as a tabular belief-space MDP that approximates the dynamics of the true MDP that we solve exactly using dynamic programming (Bellman, 1957). We note that the set B_D itself was constructed by exploring the belief-space from an initial state (uniform belief) using a ε-softmax policy associated with each $w \in \mathcal{W}$ for a given Gridworld or the entire dataset generated by participants on an experiment. This ensured that although the belief-space dynamics were approximated, this approximation was independent of the particular task or trial that was being communicated.

Pragmatic Action Interpretation

To model pragmatic action interpretation, we can extend the inverse planning process described by Equation 8 to involve inference over planning that is directed towards a literal observer's beliefs:

$$\begin{aligned} P(w \mid D, G_I, O_L, G_B) \propto & \\ \sum_{\pi} P(D \mid \pi, w)P(\pi \mid w, G_I, O_L, G_B)P(w) & \end{aligned} \tag{12}$$

Experiment 1: Modeling Details

Task Model

We model each trial as its own configuration of feature values with the same set of states, actions, transition dynamics, and discount rate, but a different environment rewards formally expressed as a utility function. To make the role of reward-based features explicit, we define a state feature function, ϕ, that maps each location state $s \in \mathcal{S}$ to a binary 5-dimensional vector where each entry corresponds to one of the colors (in order: white, yellow, orange, purple, or blue). The reward function is determined by a reward weight vector θ_w. For example, when purple and blue are dangerous, $\theta_w = [0, 10, 0, -2, -2]$. The reward for ending up in a blue state s' after taking action a in state s is determined by the feature function applied to s', $\phi(s') = [0, 0, 0, 0, 1]$, and the reward weight vector, yielding $G_I(s') = \theta_w^\top \phi(s')$. The observer starts with a uniform distribution over eight possible worlds $w \in \mathcal{W}$ and reward weights, θ_w. This corresponds to uncertainty about whether each of the orange, purple, and blue rewards are zero or -2.

Simulations

Using the task model described above, we simulated how an agent who only has instrumental utilities would act versus one who also has belief-directed utilities. For each possible world w, we calculated an instrumental demonstrator, $\pi_I(a \mid s; w)$, that serves as a model of a person who is simply doing the task. Parameter values were chosen to capture behavior that performs the task effectively with only minor deviations ($\frac{1}{\alpha} = .05$, $\varepsilon = .05$, and $\gamma = .95$). Additionally, we use these demonstrator models to define the generative model used to update a literal observer's beliefs.

For each possible world w we calculated a belief-directed demonstrator, $\pi_B(a \mid, s, b; w)$, who plans over a composite model of the task and literal observer. The model we calculated used an informativeness multiplier $\beta = 10$, and the remaining parameters were set to be the same as those of the instrumental demonstrator ($\frac{1}{\alpha_B} = .05$, $\varepsilon_B = .05$, and $\gamma_B = .95$).

For the instrumental agent, $\pi_I(a \mid s; w)$, we generated simulated trajectories by initializing it at the starting tile and then repeatedly sampling actions and transitioning to next states until it reached the goal. The same was done for the belief-directed agent, $\pi_B(a \mid, s, b; w)$, except we also initialized the observer's belief state as a uniform distribution over the eight possible reward structures and recorded the new belief state at each timestep. Each agent was simulated on each task 100 times.

Demonstrator Model-fitting

We focused on fitting belief-directed demonstrator models to each participant. To fit belief-directed demonstrators, π_B, to individual participants, we consider a space of models parameterized by seven values: The discount rate and ε-softmax values of the demonstrator's model of the observer's model of instrumental planners $(\tilde{\gamma}, \tilde{\alpha}, \tilde{\varepsilon})$; the showing discount rate and ε-softmax values of the belief-directed demonstrator $(\gamma, \alpha, \varepsilon)$; and the belief-directed reward weight (β). Since literal belief transitions are determined by how well an action distinguishes one possible world w from another, the parameters of the generative model of the inverse planner $(\tilde{\gamma}, \tilde{\alpha}, \tilde{\varepsilon})$ control how informative actions are expected to be for the observer. Meanwhile, the parameters involved in belief-directed planning $(\gamma, \alpha_{\mathrm{B}}, \varepsilon, \beta)$ reflect a communicative demonstrator's general motivation and strategy for conveying information. We searched the parameters shown in Table 1, and maximum likelihood parameter estimates are shown in Table 2.

Instrumental planning is a special case of belief-directed planning ($\beta = 0$ or $\tilde{\varepsilon} = 1.0$ or $\tilde{\alpha} \to \infty$). Thus, to assess whether belief-directed planning explains behavior in Show better than instrumental planning, we conducted likelihood-ratio tests with $\tilde{\alpha} = 1000$, $\varepsilon = 1$, and $\beta = 0$ as the null model. This makes the total difference in degrees of freedom four per model. As reported in the main text, we compared fitted instrumental planners with belief-directed planners and found that the latter better accounted for the data in Show.

Parameter	Values
$\tilde{\gamma}$	.8, .85, .9, .95, .99, .9999
$\tilde{\varepsilon}$	0.0, .025, .05, .075, .1, .125, .15, .175, .2
$\tilde{\alpha}^{-1}$	0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0
γ	.8, .85, .9, .95, .99
ε	.01, .02, .03, .04, .05, .06, .07, .08, .09, .1, .2, .3
α^{-1}	0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0
β	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25

Table 1

Experiment 1: Model-parameters evaluated using grid search.

Parameter Description		Do	Show
Discount Rate (nested)	$\tilde{\gamma}$	0.96 (0.01)	0.93 (0.01)
Random Choice (nested)	$\tilde{\varepsilon}$	0.12 (0.02)	0.09 (0.01)
Softmax Temperature (nested)	$\tilde{\alpha}^{-1}$	2.20 (0.25)	1.64 (0.25)
Belief-directed utility weight	β	2.55 (0.74)	5.31 (1.35)
Discount Rate	γ	0.93 (0.01)	0.93 (0.01)
Random Choice	ε	0.04 (0.01)	0.05 (0.01)
Softmax Temperature	α^{-1}	0.15 (0.03)	0.22 (0.04)

Table 2

Experiment 1a model parameter estimates. Means and standard errors across participants ($n = 29$ for each condition).

Experiment 2: Modeling Details

Results

Task Model

Similar to Experiment 1, each trial can be modeled as a parameterization of the transition function, $P(s' \mid s, a; w)$. We define a state feature function, ϕ that maps each tile state $s \in \mathcal{S}$ to a 6-vector where the first four entries are binary and correspond to color (white, yellow, red, green), and the last two entries correspond to the x, y coordinates of the tile. The distribution over next states given the previous state and action are defined using transformations over the different features. For example, on a strong jumper trial, $w = \texttt{Strong}$, taking the action $\uparrow$ from a green tile increments the value of the x feature by two with probability $3/4$, and by one with probability $1/4$ (assuming that the green tile is at least two tiles away from the top edge of the grid). On each trial, the observer starts with a uniform distribution over two transition functions corresponding to the green tiles being strong or weak.

Simulations

Using the above task model for each trial, we simulated an instrumental planner, π_I, and a belief-directed planner, π_B. Except for the communicative reward, which was set to $\beta = 5$ to be commensurate with the goal reward, the same parameters were used as in Experiment 1. For each trial we generated 100 trajectories, and the procedure for generating trajectories was the same as in Experiment 1.

Demonstrator Model-Fitting

Separate belief-directed planning models were fit to each participant in the two conditions, each of which had seven parameters. These were then compared with a null model in which $\frac{1}{\tilde{\alpha}} \to \infty$, $\tilde{\varepsilon} = 1$, and $\beta = 0$, which is equivalent to an instrumental planning model. Searched values are shown in Table 3, and maximum likelihood parameter estimates are shown in Table 4.

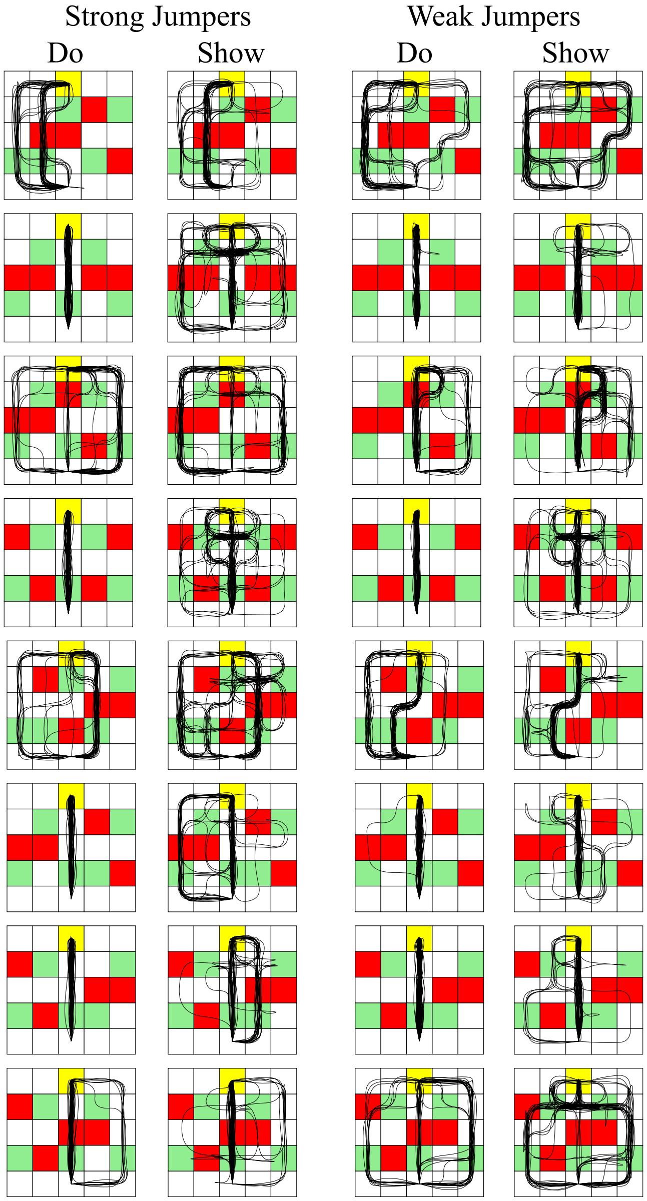

Figure 14. Experiment 2a participant trajectories by condition and trial.

Infant Observer Studies Modeling

Butler & Markman, 2012 Model Formulation

For the model, we first specify some assumptions about the observer's prior beliefs: (1) She knows the demonstrator has the goal of putting the blicket away; (2) she does not know whether blickets are magnetic; and (3) she believes that blickets are more likely to be non-magnetic than magnetic. Thus, formally, the observer starts with a distribution over two possibilities, w_{Mag} and w_{Inert}. When $w = w_{\text{Mag}}$, blickets are magnetic, and when they interact with paperclips they stick to them with a high probability, p_{Stick}. Additionally, we also assume that it is possible for the paperclips to stick to the blicket because of some alternative (unspecified) cause that is entirely independent of magnetism. This is determined by the alternative sticking probability p_{Alt}. If blickets are magnetic, then the probability of sticking is calculated with a noisy-or distribution (Pearl, 1988).

The demonstrator starts in a state where the blicket is on the table and can either put it away or put it on the paperclips. If he chooses *Put Away*, this will most likely result in Blicket Put Away, but there is a small probability of him accidentally *slipping* and the blicket landing on the paperclips ($p_{\text{Slip}} = 0.20$) before it is then put away. If he chooses *Put on Paperclips*, then it lands on the paperclips with probability 1 before being put away. Whether the paperclips and blicket

Parameter	Values
$\tilde{\gamma}$	.1, .2, .3, .4, .5, .6, .7, .75, .8, .85, .9, .95, .99
$\tilde{\varepsilon}$	0.0, .02, .06, .08, .12, .16, .18, .22, .26, .28, .32, .36, .38, .42, .46, .48
$\tilde{\alpha}^{-1}$	0.00, 0.05, .1, 0.15, .2, .25, .3, .35, .4, .45, .5, .55, .6, .65, .7, .75, .8, .85, .9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0
γ	.1, .2, .3, .4, .5, .6, .7, .75, .8, .85, .9, .95, .99
ε	0.0, .02, .06, .08, .12, .16, .18, .22, .26, .28, .32, .36, .38, .42, .46, .48
α^{-1}	0.00, 0.05, .1, 0.15, .2, .25, .3, .35, .4, .45, .5, .55, .6, .65, .7, .75, .8, .85, .9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0
β	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25

Table 3
Experiment 2a: Model-parameters searched in gridsearch.

	Do	Show
$\tilde{\gamma}$	0.58 (0.05)	0.78 (0.05)
$\tilde{\varepsilon}$	0.25 (0.03)	0.24 (0.03)
$\tilde{\alpha}^{-1}$	0.83 (0.27)	1.11 (0.28)
β	1.49 (0.33)	5.68 (0.78)
γ	0.84 (0.04)	0.76 (0.04)
ε	0.03 (0.01)	0.18 (0.02)
α^{-1}	0.05 (0.01)	0.08 (0.01)

Table 4
Experiment 2a model parameter estimates. Means and standard errors across participants (n_{Do} = 39, n_{Show} = 41).

stick together depends on whether blickets are magnetic or inert, as described in the previous paragraph. The instrumental utilities are +1 for putting the blicket away and -0.1 for each action taken (e.g., putting it on the paperclips and then putting it away is 2 steps).

This formulation of the task allows us to distinguish between the blicket *accidentally* landing on the paperclips, which occurs in the Accidental condition, and the blicket *intentionally* landing on the paperclips, which occurs in both the Intentional and Communicative conditions (Figure 11a). The accidental demonstration can be modeled as the sequence where the demonstrator first takes the action *Put Away*, but then slips and lands on the Paperclips Attached state before ending on the Blicket Put Away state. In contrast, the intentional/communicative demonstrations directly place the blicket on the paperclips by selecting *Put On Paperclips*, having them attach, and then putting it away.

All demonstrator models select actions using a softmax policy with $\alpha = 0.2$ (there is no random choice; $\varepsilon = 0.0$). Although Butler and Markman (2012) report two measures of exploration on a different task, this is primarily in order to assess the strength of the inference about whether blickets are magnetic. Thus, we report the probabilities calculated by our model directly rather than make any assumptions about how these relate to exploratory behavior. As shown in Figure 15, the equivalance of the Intentional and Accidental conditions as well as the higher belief in blicket magnetism in the Communicative condition are consistent across a range of parameters.

Hernik & Csibra, 2015 Model Formulation

Although the studies in question involve multiple counterbalanced training trials, in order to understand how the key findings relate to our account it suffices to explore the inferences our models make after observing a single training trial. Specifically, we model a trial in which the banana's initial state is Unpeeled and its final state is either Peeled or Unpeeled. Additionally, we make the following assumptions

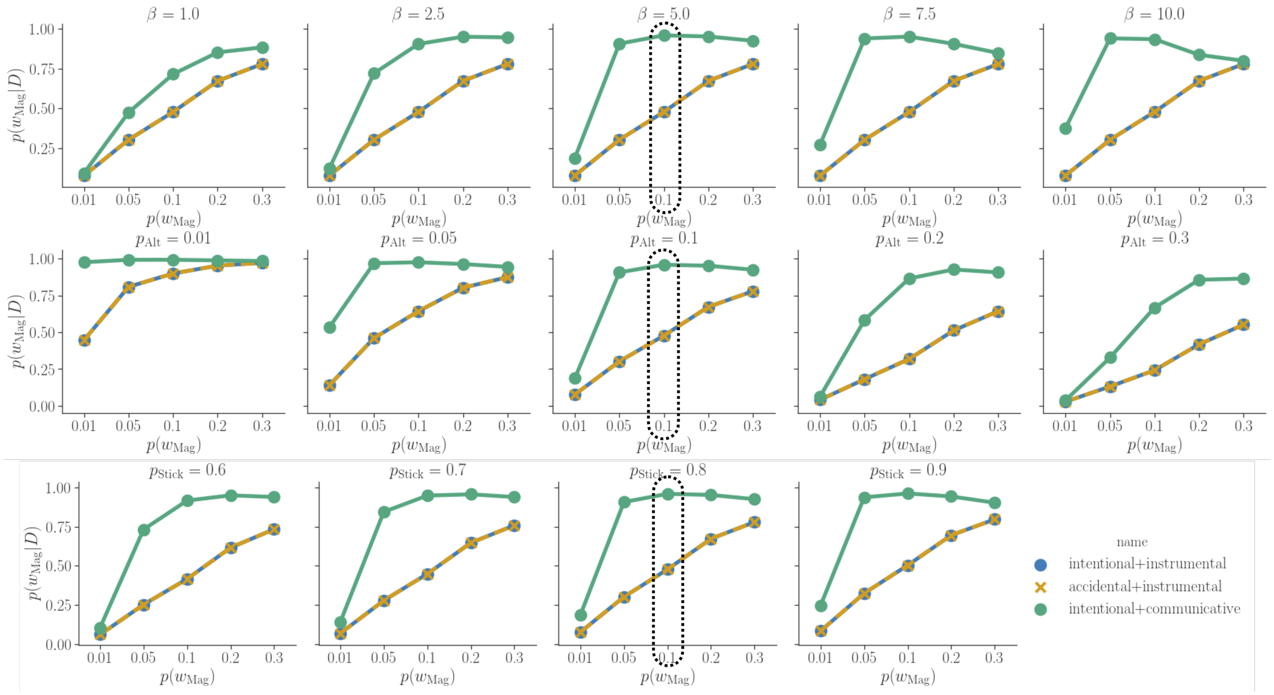

Figure 15. Behavior of observer models for Butler & Markman, 2012 for a range of parameter values. $p(w_{Mag}$ is the prior probability of blicket magnetism; $p(w_{Mag} \mid D)$ is the posterior belief in blicket magnetism having observed the demonstration; p_{Alt} is the alternative cause probability; p_{Stick} is the the probability of paperclips sticking given blickets are magnetic; β is the teaching weight bias. The points enclosed in the dotted line corresponds to the parameters reported in the main text as a point of reference. For a range of teacher weights (top row), alternative cause probabilities (middle row), and magnetic strength values (bottom row), the Intentional and Accidental conditions are equal while the Communicative condition is substantially higher, mirroring the general pattern of results found in the study. For all simulations, the planning model was held constant with random choice, $\varepsilon = 0.0$; and softmax choice probability, $\alpha = 0.2$.

about observer prior beliefs: (1) There is a background probability that the objects will change independent of tool use or effectiveness, and (2) arbitrary tools and arbitrary objects do not usually causally interact. We note that although the participants never see the banana changed independently of the tool, they must be able to represent the possibility that the tool was *not* the cause of the banana's transformation. Thus, although it does not need to be exactly specified, there must be some alternative cause of the transformation which is why we assume there is some non-zero probability that the objects will change independent of tool use. Formally, we assume that a background probability of objects changing due to an alternative cause, p_{Alt}; that there are two relevant possible worlds w where either the banana is a peeler (w_{Peeler}) or not (w_{Inert}); and that the initial probability of the tool being a banana peeler is low (i.e. $b(w_{\mathrm{Peeler}}) < .5$).

The demonstrator starts in the Unpeeled state and can choose either *Do Nothing* or *Use Tool*. If he chooses *Do Nothing*, then regardless of whether the tool is a banana peeler or not the state transitions to Peeled or Unpeeled according to the background probability. On the other hand, if he chooses *Use Tool*, then the probability of transitioning depends on the specific world. If the world is w_{Peeler}, then it will transition to Unpeeled based on a combination of the background transition probability and the tool's effectiveness ($\theta_{Effectiveness}$). Specifically, we assume that these two combine in a "noisy-or" manner where the effect occurs if either cause (or both) are activated (Pearl, 1988). Additionally, we assume a small step-cost of using the tool ($-.1$) and that there is a reward for peeling the banana ($+1$). If the tool does not have the function of being a banana peeler and true world is w_{Inert}, then *Use Tool* has the same transition probabilities as *Do Nothing*. The different transition and utility functions are visualized in Figure 12b.

A linking function is required to connect the model outputs to the measure reported in the experiments. We can simulate the violation of expectation measure by calculating the *surprisal* (the negative log probability) of a congruent or incongruent trial given a model's posterior distribution. We can then use that distribution to calculate how surprised the model would be to see the congruent or incongruent test trials, where the actions are assumed to be taken instrumentally.

All demonstrator models select actions with a softmax policy ($\alpha = 0.2$). Figure 16 shows the results when parameterically varying the background probability (p_{Alt}), the tool efficacy ($\theta_{Efficacy}$), and the teaching weight (β). Overall, we see that the amplification of inferences about the peeler tool increases in the communicative conditions consistently, but

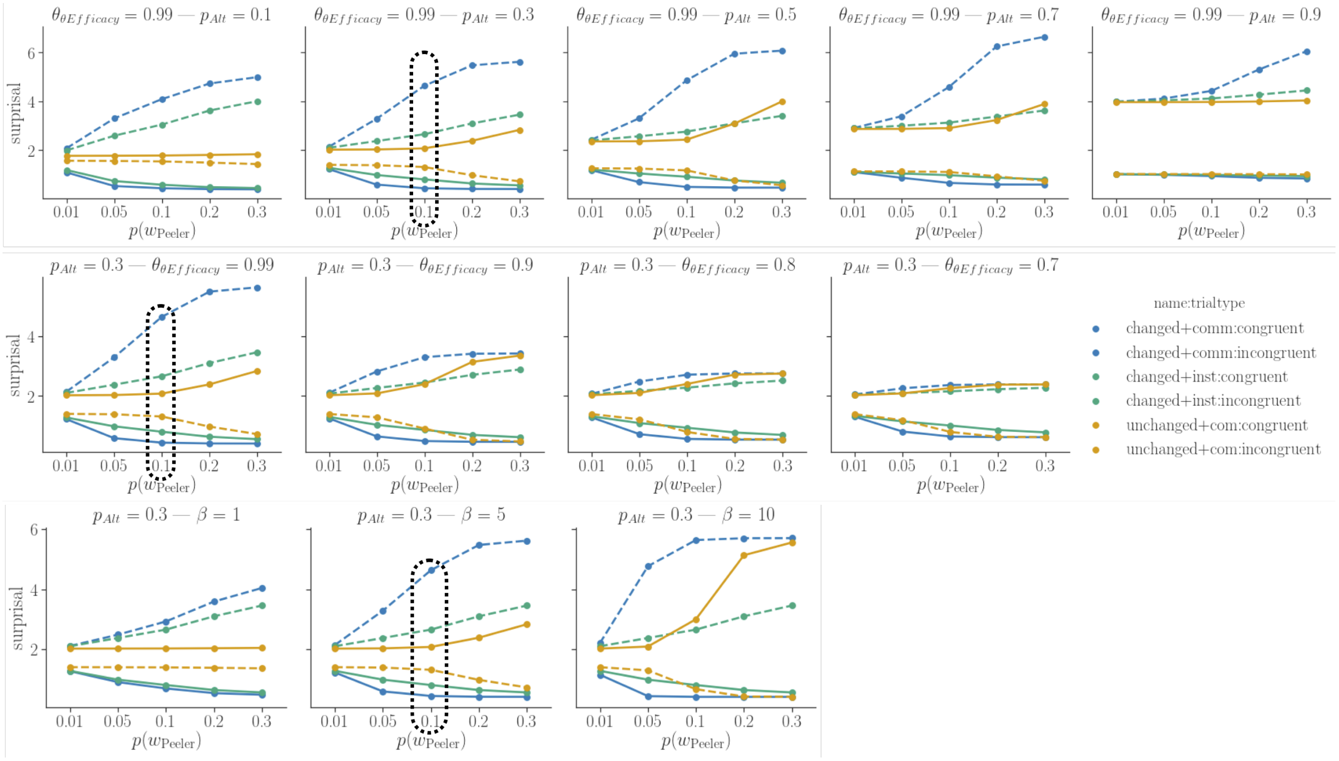

Figure 16. Behavior of pragmatic observer model for Hernik & Csibra, 2015 for a range of parameter values. $p(w_{\text{Peeler}})$ is the prior probability that the tool is a peeler; p_{Alt} is the probability of an alternative (unspecified) cause of the banana changing; $\theta_{Efficacy}$ is the probability that the banana changes if the tool is in fact a peeler; and β is the teaching weight. The points enclosed in the dotted line indicate the set of values that are plotted in the main text to provide a point of reference. Colors correspond to the study modeled, solid lines are the congruent trials, and dotted lines are the incongruent trials. Across all parameterizations, the communicative trials in which the banana changed lead to stronger versions of the inferences made in the non-communicative trials (green versus blue lines). The inferences made in the communicative trial where the banana did not change is less consistent across parameterizations (yellow lines), indicating that the inferred communicative intent in such an ambiguous situation is sensitive to background beliefs. For all simulations, the planning model was held constant with random choice, $\varepsilon = 0.0$ and softmax choice probability, $\alpha = 0.2$.

that the inferences when the tool is communicatively presented *without* any change are more sensitive to prior beliefs about the peeler and the probability of alternative causes.

Király et al., 2013 Model Formulation

We can formalize the experiment in our modeling framework. Specifically, we begin by specifying the following two assumptions about the observer's prior: (1) Whether the box lights up when touched is initially unknown to observer, and (2) a demonstrator is more likely to have using their hands as a subgoal rather than using their head. Formally, beliefs about the box being a light is represented with a distribution over a binary variable $p(w_{\text{Box-is-Light}})$. Subgoals are represented as action priors (Wingate, Goodman, Roy, Kaelbling, & Tenenbaum, 2011) that operate as a bias over different actions in the following manner: The function $\bar{A}$ assigns a prior probability to each action, and $\bar{A}(a) = 1$. Uncertainty about subgoals is then represented as a distribution over different action priors, $\bar{A}$. The observer considers two possible action priors, $\bar{A}_{Hand}$ and $\bar{A}_{Head}$, paramterized by an action bias strength $\theta_{\bar{A}} \in [0, 1]$, where $\bar{A}_{\text{Action}}(a) = \theta_{\bar{A}}$ if a matches Action and $\bar{A}_{\text{Action}}(a) = 1 - \theta_{\bar{A}}$ if not. Note that we use a softmax action rule, so the action prior can be incorporated into the Q-value as $\log \bar{A}(a)$ (see Equation 13 below). To summarize, the learner's prior requires specifying three parameters: the distributions $p(w_{\text{Box-is-Light}})$ and $p(\bar{A})$, and the action bias strength $\theta_{\bar{A}}$.

The experimental setup itself can be modeled as a demonstrator who begins in a state s that has variables with values, s_{Box} = `Unlit` and $s_{Hands} \in$ {`Free`, `Occupied`}. If s_{Hands} = `Free`, then they have three actions available, $\mathcal{A}(s)$ = {`Do Nothing`, `Use Hand`, `Use Head`}, but if s_{Hands} = `Occupied`, then $\mathcal{A}(s)$ = {`Do Nothing`, `Use Head`}. That is, they can only use their head if their hands are occupied. Taking an action potentially modifies the state such that s'_{Box} = `Lit` or s'_{Box} = `Unlit` depending on the value of $w_{\text{Box-is-Light}}$. The demonstrator plans and selects actions based on the expected utility of an action from a state, taking into account instrumental goals (G_I), action biases ($\bar{A}$), and

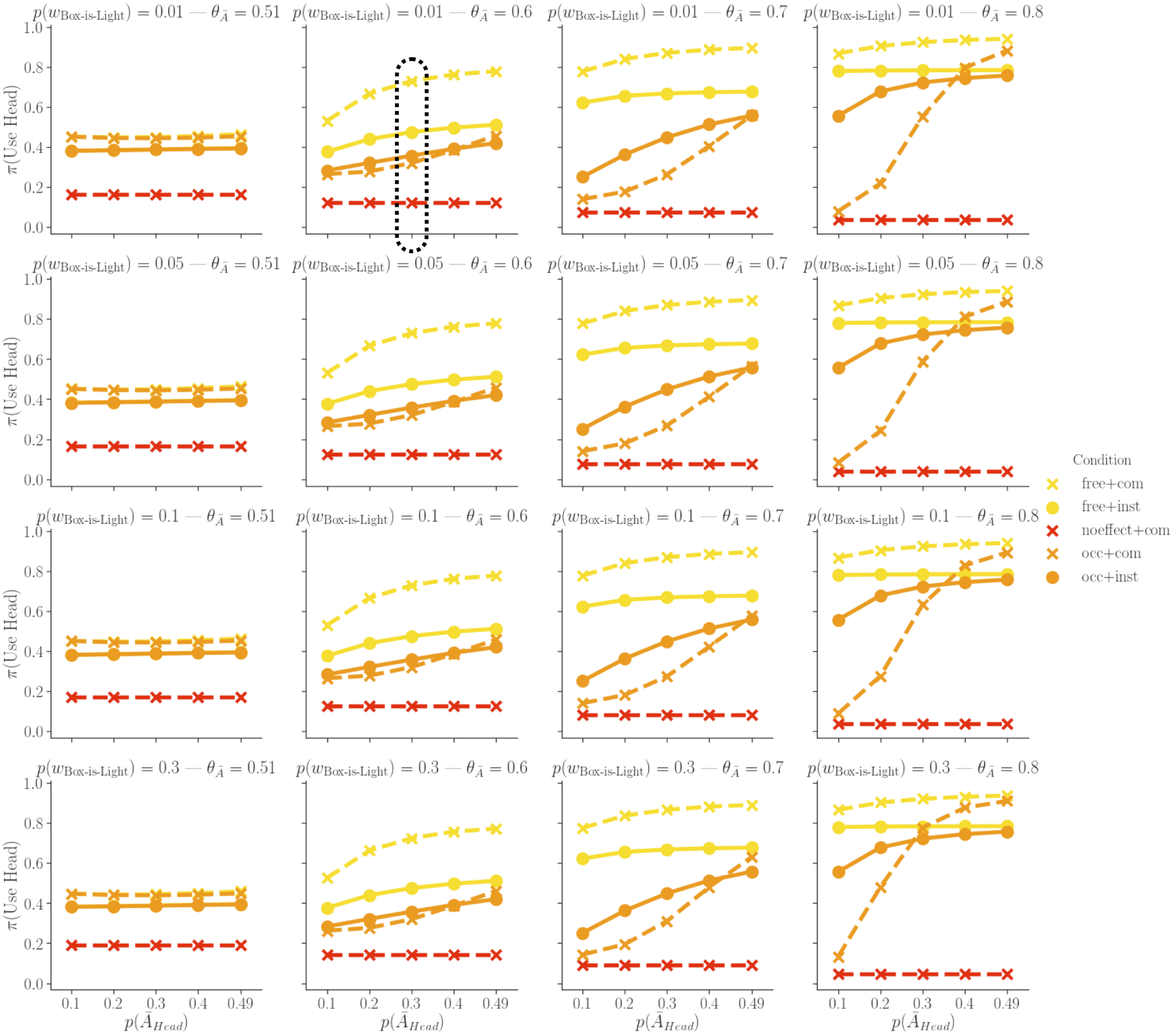

Figure 17. Behavior of pragmatic observer model for Király et al., 2013 for a range of parameter values. $p(w_{Box-is-Light})$ is the prior probability that the box lights up; $\theta_{\bar{A}}$ is the subgoal bias strength; π(Use Head) is the probability that the observer model, having observed a demonstration in a context, imitates the action *Use Head*; $p(\bar{A}_{Head})$ is the prior probability that using one's head and not one's hand is a subgoal, given turning on the light is a goal. The points enclosed in the dotted line correspond to the parameters reported in the main text as a point of reference. The general pattern of results that communicative demonstrations (dotted lines) lead to more extreme imitation of using one's head or not holds across a range of parameters as long as the subgoal bias ($\bar{A}$) is sufficiently greater than .5, and the novelty of the head action (i.e. $1 - p$(Use Hand Subgoal)) is high. The red lines correspond to the *No Effect* condition from Experiment 2 reported by Király et al., 2013 and are consistently lower than all the other conditions. For all simulations, planning decision rule was held constant with the teaching weight, $\beta = 1$; random choice, $\varepsilon = 0.0$; and softmax choice probability, $\alpha = 0.2$.

communicative goals (G_C):

$$\begin{aligned} Q(a, s, b; w, \bar{A}) = & \\ \sum_{s', b'} P(s' \mid s, a; w) O_L(b' \mid s, b, a) \Big[& \\ & G_I(s'; w) \\ & + \log \bar{A}(a) \\ & + \beta G_C(b', b; w) \Big] \end{aligned} \tag{13}$$

In our implementation, the reward associated with turning the light on was always 1. Additionally, the demonstrator's action selection rule always had a softmax parameter, $\alpha^{-1} = 0.2$ and no random choice ($\varepsilon = 0.0$).

Since Király et al., 2013 operationalized social learning by measuring the rate of head-action imitation, we need a linking function from resulting posterior beliefs (i.e., $b(w \mid s, a, s')$) to behavior (i.e., $\pi(a \mid s, b')$). To model how the infant observers would act after having observed a demonstration, we calculated the policy that is optimal in expectation based on the resulting observer belief b', and report the softmax policy probabilities ($\frac{1}{\alpha} = 2.5$) when $s_{Hands} = \texttt{Free}$ and $s_{Box} = \texttt{Unlit}$.

Using this setup, we modeled five of the experimental conditions reported by Király et al., 2013: the Communicative/Instrumental x Hands Occupied/Hands Free conditions reported in Experiment 1, and the No Effect condition in Experiment 2, in which the demonstrator ostensively cued the participant before using their head to *try* and turn on the box without it turning on. Figure 17 shows the outputs of the model for the different conditions over a range of parameterizations of prior beliefs. In general, we find that the model captures the qualitative patterns reported in the original studies.