Supplementary Materials for “Communication in Action: Planning and Interpreting Communicative Demonstrations”

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:

$$
\begin{aligned}
&Q(a, s; w) = \\
&\quad \sum_{s'} P(s' \mid s, a; w) \left[G_I(s, a, s') + \gamma \max_{a'} Q(a', s'; w) \right],
\end{aligned} \tag{1}
$$

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:

$$
\pi(a \mid s; w) = (1 - \varepsilon) \frac{e^{\alpha Q(s,a;w)}}{\mathcal{Z}(s; w)} + \frac{\varepsilon}{|\mathcal{A}(s)|}, \tag{2}
$$

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:

$$
P(D \mid \pi, w) = \prod_{t=0}^{T} \pi(a_t \mid s_t; w) P(s_{t+1} \mid s_t, a_t; w) \tag{3}
$$

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, Saxe, & Tenenbaum, 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 3:

$$\begin{aligned} P(w \mid D, G_I) &\propto P(D \mid w, G_I)P(w) \\ &= \sum_{\pi} P(D \mid \pi, w)P(\pi \mid w, G_I)P(w) \end{aligned} \tag{4}$$

As we discuss in the main text, this process of inverse planning can be used to define a *literal observer model* $\mathcal{O}_L$ by associating beliefs b with probability distributions that are updated according to Equation 4. Specifically, the one-step literal observer belief state updates upon observing a state, action, and next-state are given by:

$$\begin{aligned} &\mathcal{O}_L(b' \mid s, b, a, s') \\ &\qquad = \begin{cases} 1 & \text{if for all } w, \\ & b'(w) \propto \pi(a \mid s; w)P(s' \mid s, a; w)b(w) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

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, Brunskill, Griffiths, & Shafto, 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, $\mathcal{O}_L$, and belief-directed goals, G_B. Formally, the Q-values for a belief-directed agent are:

$$Q(a, s, b; w) = \sum_{s', b'} P(s' \mid s, a; w)\mathcal{O}_L(b' \mid s, b, a)\Big[G_I(s, a, s'; w) + \beta G_B(b', b; w) + \gamma \max_{a'} Q(a', s', b'; w)\Big] \tag{6}$$

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:

$$G_B(b', b; w) = b'(w) - b(w). \tag{7}$$

Given the Q-values over joint ground and belief states (Equation 6), 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 $b_D \in B_D$, $\hat{P}(s', b'_D \mid a, s, b_D) = \sum_{b'} P(s' \mid s, a; w)\mathcal{O}_L(b' \mid b_D, s, a, s')NN(b', b'_D)$, 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 4 to involve inference over planning that is directed towards a literal observer's beliefs:

$$\begin{aligned} &P(w \mid D, G_I, \mathcal{O}_L, G_B) \propto \\ &\qquad \sum_{\pi} P(D \mid \pi, w) P(\pi \mid w, G_I, \mathcal{O}_L, G_B) P(w) \end{aligned} \tag{8}$$

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

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.

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, $\pi_B(a\ |, s, b; w)$, 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, $\pi_B(a\ |, s, b; w)$, 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

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).

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.

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.

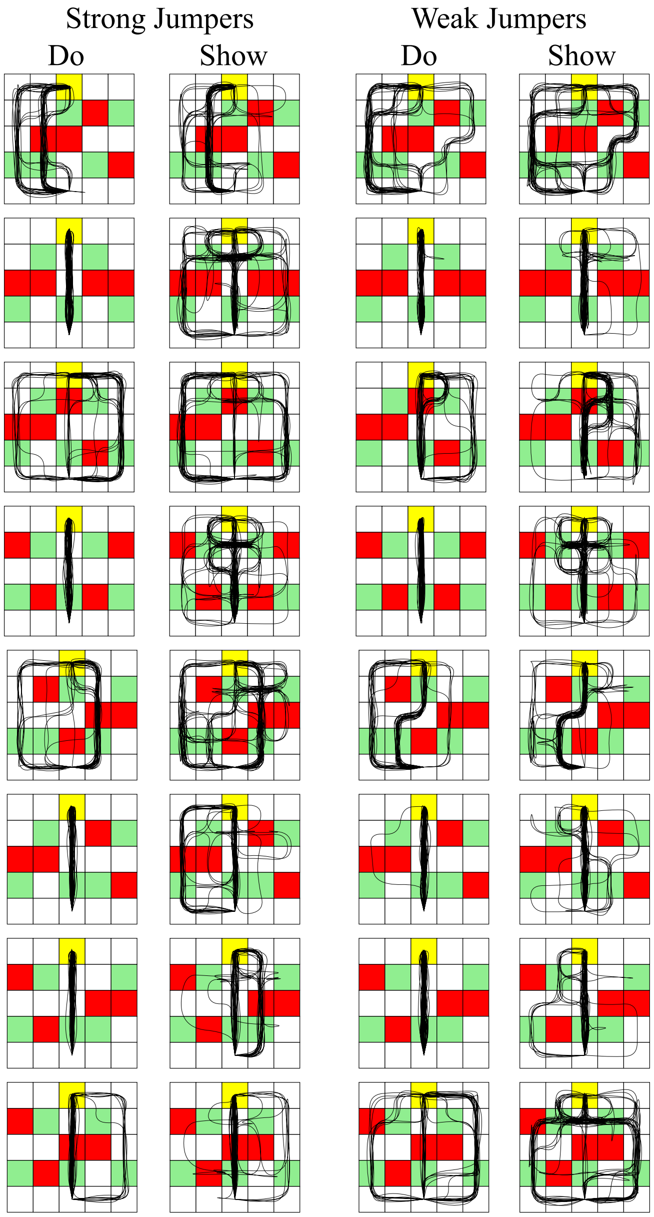

Figure 1. Experiment 2a participant trajectories by condition and trial.

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

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.

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

	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 ($n_{Do} = 39$, $n_{Show} = 41$).

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 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 (see main text). 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

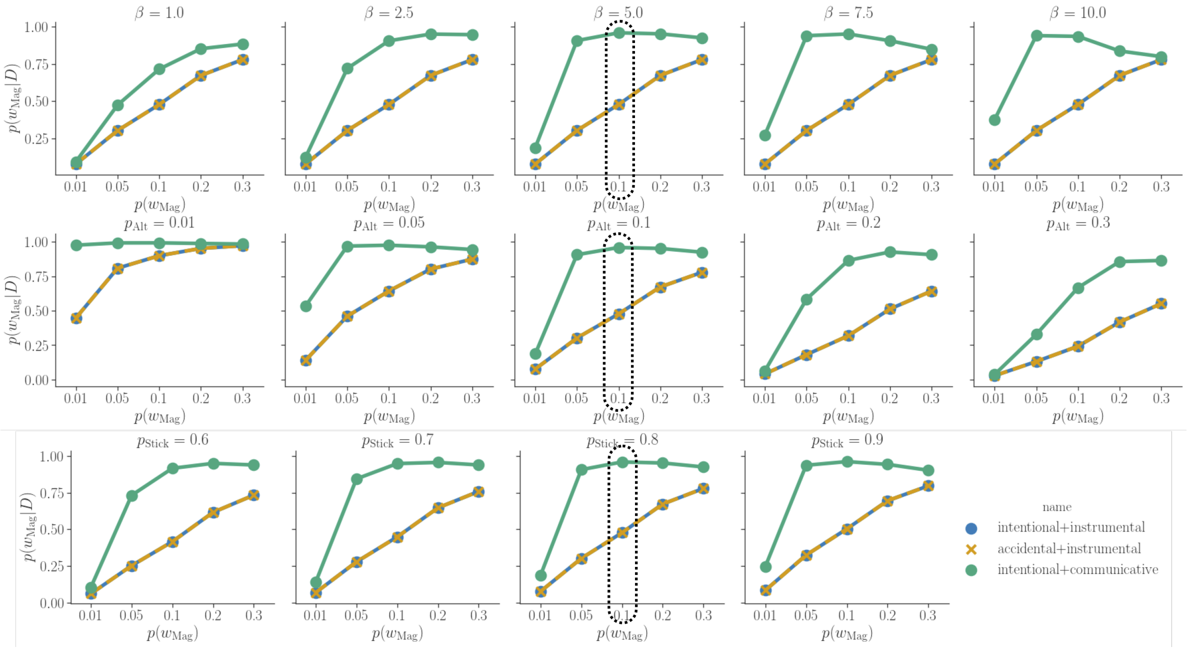

Figure 2. Behavior of observer models for Butler & Markman, 2012 for a range of parameter values. $p(w_{Mag}$ is the prior probability of blicket magnetism; $p(w_{Mag} \mid D)$ is the posterior belief in blicket magnetism having observed the demonstration; p_{Alt} is the alternative cause probability; 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, $\alpha = 0.2$.

directly rather than make any assumptions about how these relate to exploratory behavior. As shown in Figure 2, 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.

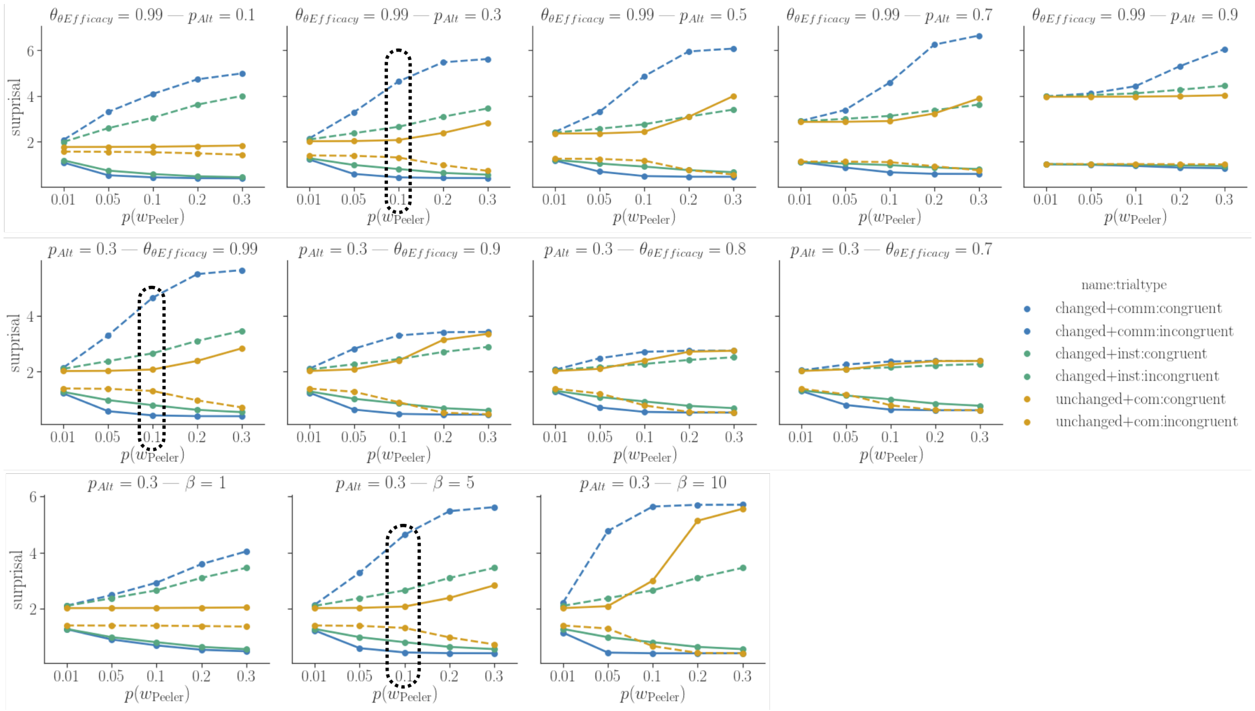

Figure 3. Behavior of pragmatic observer model for Hernik & Csibra, 2015 for a range of parameter values. $p(w_{\text{Peeler}})$ is the prior probability that the tool is a peeler; p_{Alt} is the probability of an alternative (unspecified) cause of the banana changing; $\theta_{Efficacy}$ is the probability 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, $\alpha = 0.2$.

Additionally, we make the following assumptions 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. $b(w_{\text{Peeler}}) < .5$).

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 the main text.

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 3 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 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, Csibra, & Gergely, 2013 Model Formulation

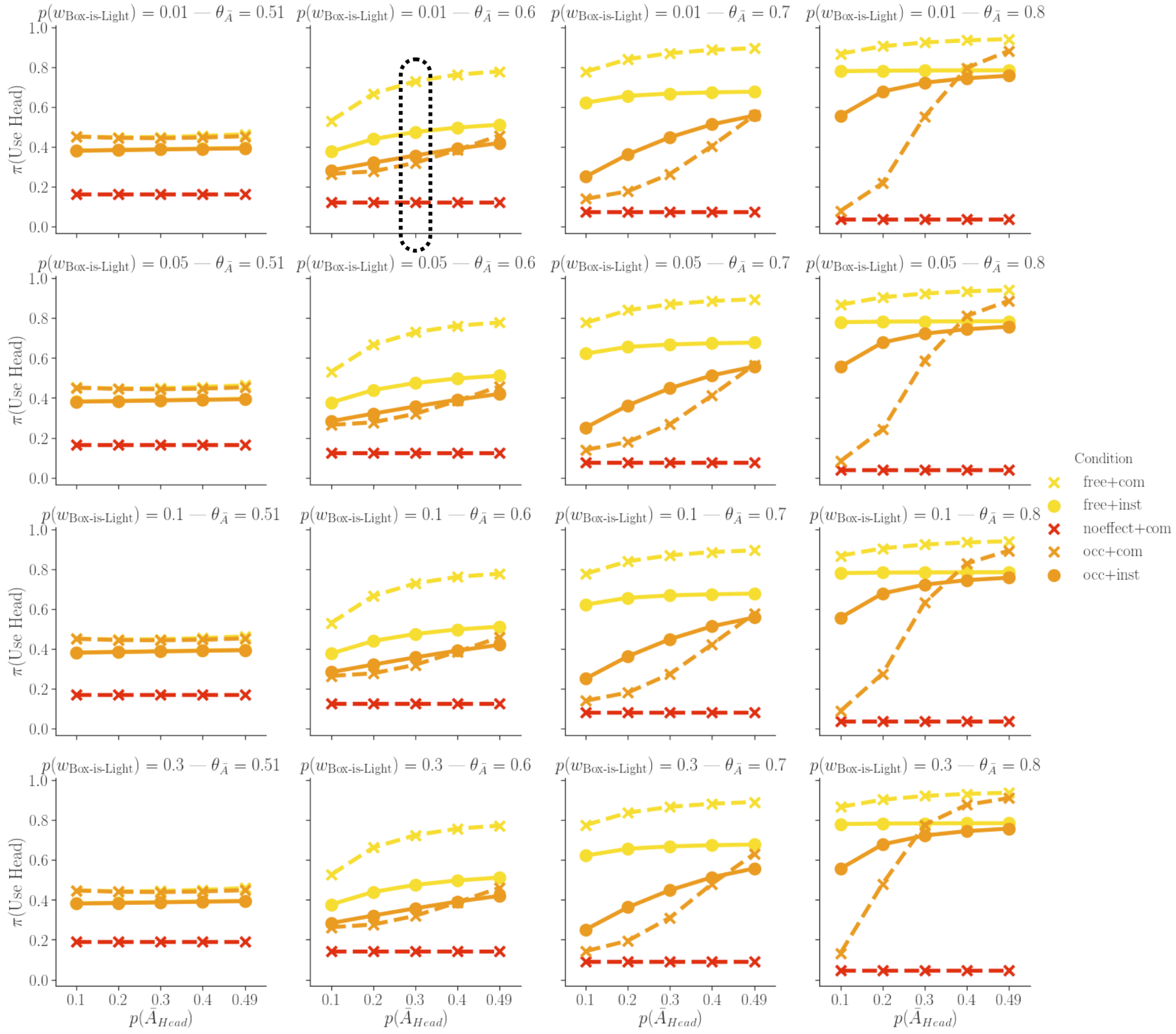

Figure 4. Behavior of pragmatic observer model for Király et al., 2013 for a range of parameter values. $p(w_{Box-is-Light})$ is the prior probability that the box lights up; $\theta_{\bar{A}}$ is the subgoal bias strength; π(Use Head) is the probability that the observer model, having observed a demonstration 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, $\alpha = 0.2$.

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 9 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, $s_{Box} = \texttt{Unlit}$ and $s_{Hands} \in \{\texttt{Free}, \texttt{Occupied}\}$. If $s_{Hands} = \texttt{Free}$, then they have three actions available, $\mathcal{A}(s) = \{\texttt{Do Nothing}, \texttt{Use Hand}, \texttt{Use Head}\}$, but if $s_{Hands} = \texttt{Occupied}$, then $\mathcal{A}(s) = \{\texttt{Do Nothing}, \texttt{Use Head}\}$. That is, they can only use their head if their hands are occupied. Taking an action potentially modifies the state such that $s'_{Box} = \texttt{Lit}$ or $s'_{Box} = \texttt{Unlit}$ 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 communicative goals (G_C):

$$
\begin{aligned}
&Q(a, s, b; w, \bar{A}) = \\
&\qquad \sum_{s', b'} P(s' \mid s, a; w) \mathcal{O}_L(b' \mid s, b, a) \Big[G_I(s'; w) + \log \bar{A}(a) + \beta G_C(b', b; w) \Big]
\end{aligned} \tag{9}
$$

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 4 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.