Developmental Science

Developmental Science 16:2 (2013), pp 209–226 DOI: 10.1111/desc.12017

PAPER

The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model

J. Kiley Hamlin,[1] Tomer Ullman,[2] Josh Tenenbaum,[2] Noah Goodman[3] and Chris Baker[2]

1. *Department of Psychology, University of British Columbia, Canada*
2. *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA*
3. *Department of Psychology, Stanford University, USA*

Abstract

Evaluating individuals based on their pro- and anti-social behaviors is fundamental to successful human interaction. Recent research suggests that even preverbal infants engage in social evaluation; however, it remains an open question whether infants' judgments are driven uniquely by an analysis of the mental states that motivate others' helpful and unhelpful actions, or whether non-mentalistic inferences are at play. Here we present evidence from 10-month-olds, motivated and supported by a Bayesian computational model, for mentalistic social evaluation in the first year of life.A video abstract of this article can be viewed at http://youtu.be/rD_Ry5oqCYE

Introduction

A growing body of evidence suggests that foundational aspects of social evaluation are present extremely early in human ontogeny. Infants in the first year of life distinguish positive and negative social acts (Premack & Premack, 1997); react differentially to those who direct helpful versus harmful intentions toward them, despite superficially similar behaviors and identical outcomes (e.g. Behne, Carpenter & Tomasello, 2005; Marsh, Stavropoulos, Neinhuis & Legerstee, 2010; see also Dunfield & Kuhlmeier, 2010); and expect others to prefer those who have helped versus hindered them (e.g. Kuhlmeier, Wynn & Bloom, 2003). In addition, infants themselves positively evaluate those who have helped versus harmed unknown third parties (e.g. Hamlin & Wynn, 2011; Hamlin, Wynn & Bloom, 2007, 2010; see also Geraci & Surian, 2011). In one study (Hamlin *et al.*, 2007), 6- and 10-month-olds saw a climber try but fail to climb a hill; it was then bumped up and down the hill by two additional characters. When given the choice between the helpful and unhelpful characters, infants at both ages preferred the helpful one, suggesting that they evaluated the individuals based on their third-party social acts. Further studies have demonstrated that social evaluations occur in infants as young as 3 months, and are applied to multiple kinds of agents (wooden shapes with eyes, animal puppets, computer animations). They are engaged during a variety of goal facilitation/prevention situations in addition to hill climbing, including opening/closing a box lid, and giving/taking a ball (Hamlin & Wynn, 2011).

But how, specifically, are infants reasoning about social actions in order to evaluate helpful and harmful individuals? Opposing accounts offer different answers. According to the mentalistic account, infants have a theory of intuitive psychology that is sensitive to the mental states of other agents, and can reason over them. Indeed, a large literature suggests that infants in the first year of life grasp the influence of mental states on others' acts, privileging goals and preferences over physical aspects of behaviors (e.g. Biro & Leslie, 2007; Csibra & Gergely, 2003; Luo, 2011; Luo & Baillargeon, 2005; Woodward, 1998, 1999). They do so even when physical information is inconsistent with goal-states, as is the case when goals go unfulfilled (e.g. Brandone & Wellman, 2009; Hamlin, Newman & Wynn, 2009), or an agent's representation of a scene is less complete than their own (e.g. Luo & Baillargeon, 2007; Luo & Johnson, 2009; Tomasello & Haberl, 2003). This mentalistic account

Address for correspondence: J. Kiley Hamlin, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC V6T1Z4, Canada; e-mail: Kiley.hamlin@psych.ubc.ca

claims that infants represent the goals and beliefs of agents who need help, as well as the goals and beliefs of individuals who help and hinder those needy agents, and use this information to select helpers over hinderers. In particular, the account suggests that infants represent helpfulness/harmfulness as a type of goal or desire on the part of helpers/hinderers: it is the helper's goal to bring about the goal of the needy agent, whatever it may be; it is the hinderer's goal to prevent the goal of the needy agent, whatever it may be.

The computations required for this mentalistic account are not trivial: specificially, they involve second-order mental-state representations (the goal of one agent depends on the goal of another agent). While recent research suggests that infants are capable of sophisticated mental-state reasoning (e.g. Onishi & Baillargeon, 2005; Southgate, Senju & Csibra, 2007; Surian, Caldi & Sperber, 2007), these studies have utilized older infants, at least in the second year of life, and have not escaped criticism (e.g. Ruffman & Perner, 2005). Indeed, a number of lower-level accounts of infants' social evaluations are plausible, and need to be ruled out before a high-level, mentalistic account is adopted.

According to a low-level cue-based account, infants are not carrying out complicated reasoning processes over mental states, but rather using features present in the displays to evaluate the observed behaviors. For example, perhaps infants prefer helpers to hinderers in the hill scenario because they prefer those who push things uphill versus downhill, or in the box scenario (of Hamlin & Wynn, 2011) because they prefer those who open versus close boxes. More generally, infants might prefer agents that facilitate former action sequences – e.g. if a Protagonist moved up, it's good to push it up. Such an account would side-step the difficulty of mental state reasoning in formal definitions of 'helping' and 'hindering'. Importantly, however, all previously published infant evaluation studies have included controls in which physically congruent behaviors were directed at non-social entities (that presumably cannot possess goal states); infants did not show preferences in these conditions (e.g. Hamlin *et al.*, 2007, 2010; Hamlin & Wynn, 2011). Yet, the possibility remains that low-level behavioral cues are activated specifically in situations involving social entities.

Several mid-level accounts reside between a cue-based account and a fully mentalistic one. For example, infants might engage in first-order goal attribution, but *not* in second-order goal attribution. That is, infants could infer the goal of the needy agent only, and positively evaluate individuals who complete that goal, and negatively evaluate individuals who prevent it. This could occur without infants holding any belief that helpers have the goal to help or hinderers have the goal to hinder; instead, infants could analyze only the completion, or not, of the goal of the needy agent. Indeed, if infants analyze the likelihood that individuals might help *them* in the future, then evaluation based on tendencies to bring about others' goals, rather than intentions to do so, might be a perfectly useful strategy. In particular, this account predicts previous failures of infants to evaluate those who direct their behaviors toward inanimates: inanimates presumably do not activate infants' first-order goal attributions (e.g. Woodward, 1998).

A more complex mid-level account might claim that infants engage in second-order *goal* attribution, but not in second-order *knowledge/belief* attribution. That is, infants might represent the goal of the needy agent, and then assume that the helper and the hinderer also represent this goal. This assumption does not require a representation of the helper's and hinderer's knowledge or beliefs about the agent's goal (in particular, that they could be incomplete or false); instead, evaluation could be based simply on assuming that everyone represents the goal as the infant him or herself does. For example, in the hill scenario, infants might assume that just as they know the first agent tried to get up the hill, all future helping and hindering agents share this knowledge. Thus, while a second-order goal representation is required, this account still lacks requirements for a fully mentalistic account of infants' social evaluations, one that incorporates helpers' and hinderers' representational mental contents.

Importantly, the preceding accounts have parallels in computational modeling approaches, which are accordingly at odds with one another. Bottom-up perceptual models categorize actions as 'helping' or 'hindering' based on low-level visual features present in the observed display (see Scholl & Tremoulet, 2000; Blythe, Todd & Miller, 1999; Gao, Newman & Scholl, 2009). Opposed to these cue- or end-state-based models are more mentalistic computational models of social evaluations based on an Inverse Planning approach and an intuitive psychology viewpoint; these have recently been introduced to explain the social inferences of adults and children (e.g. Baker, Saxe & Tenenbaum, 2009; Baker, Goodman & Tenenbaum, 2008; Ullman, Baker, Macindoe, Evans, Goodman & Tenenbaum, 2010; Ullman, Macindoe, Baker, Evans, Goodman & Tenenbaum, in preparation). Inverse Planning assumes that the child or observer represents agents as rational planners, who choose actions in order to achieve desired goals, subject to their beliefs. This planning process can then be inverted to infer unobserved beliefs and goals from observed actions. These models have been used to

explain studies on infant first-order goal inference such as those of Gergely and Csibra (Baker *et al*., 2008, 2009), as well as social-goal inferences such as helping and hindering (Ullman *et al*., 2010, in preparation). Importantly, the Inverse Planning approach accurately predicts the quantitative judgments that adults make in such scenarios.

Despite these successes with adults, it is possible that Inverse Planning models do not reflect how preverbal infants evaluate third-party helping and hindering. Furthermore, these adult studies do not distinguish between second-order goal analysis and second-order knowledge or belief analysis; different Inverse Planning models could in fact span the range from high-level accounts of social evaluation based on full second-order mental representations, to mid-level accounts based on partial second-order representations, to first-order representations. Thus, the current study was designed to distinguish between these fundamentally different accounts of infant social evaluation, guided by a computational modeling framework sufficient to express clearly the range of more or less sophisticated mentalistic accounts.

In order to test these accounts, it is important to note that in all previous studies of infant social evaluation, the actions judged to be socially different were also perceptually distinct. A crucial aspect of the Inverse Planning approach is that it can account for different social judgments about the *exact same actions*, depending on the beliefs, knowledge, and goals of the agents involved (Ullman *et al*., 2010); a cue-based approach cannot explain how a perceptually indistinguishable action is judged differently. In the current study, we set up a design similar to the one used in previous tests of social evaluation (Hamlin *et al*., 2007; Hamlin & Wynn, 2011), but varied *only* the potential mental states of the agents involved – the preference of the needy agent, and knowledge of that preference by the intervening agents. The actions of the intervening agents were identical.

We compared three experimental conditions. In the first phase of all conditions, infants observed a Protagonist puppet (Lion) repeatedly grasp an object through one of two small openings in a wall, establishing a consistent tendency to grasp that object. In a second phase, the Protagonist lost its access to the toy it was grasping when doors were placed in both openings; it attempted unsuccessfully to jump over the center of the wall. In a third phase, two 'Door-Lifter' puppets (Elephants) alternately lifted each door and allowed the Protagonist to move through the opening and grasp the object behind it; one Lifter always gave the Lion access to the object that it had grasped before, the other Lifter always gave the Lion access to an object it had not previously grasped (see Figure 1).

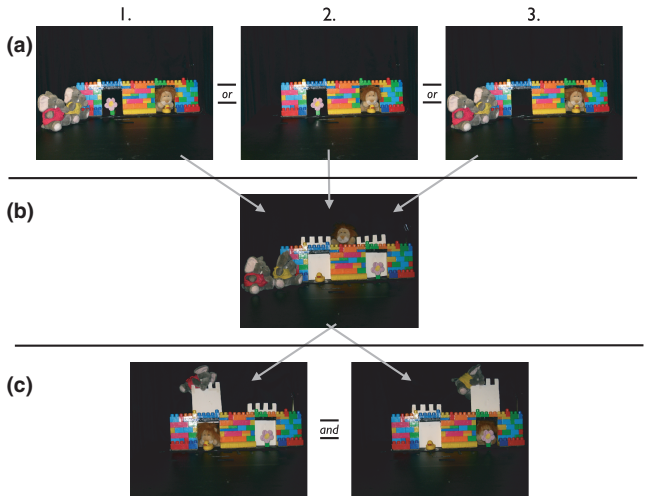

Figure 1 *Stimuli presented to infants. Row A depicts the Familiarization events shown to infants in (1) the Preference-Knowledge, (2) Preference-Ignorance, and (3) NoPreference-Knowledge conditions. Row B depicts the Baseline event shown to all infants. Row C depicts the Door-Lifting events shown to all infants.*

Across the three conditions, we varied the following:

1 Whether or not infants could attribute to the Protagonist a *preference* for the grasped toy, by varying whether there was a second object present during the first phase. Observing an agent repeatedly grasping one object (e.g. a toy duck) in the presence of another (e.g. a flower) suggests that he has a preference for the grasped object. However, observing the same agent grasping a duck while not in the presence of a flower gives observers no information about his relative preference for ducks versus flowers. A series of studies by Luo and colleagues suggests that, rather than expecting individuals to always approach objects that they have approached before, infants selectively make preference attributions based on the presence or absence of alternative target objects (e.g. Luo, 2011; Luo & Baillargeon, 2005, 2007; Luo & Johnson, 2009; see also Kushnir, Xu & Wellman, 2010). In the current studies, infants in conditions 1 and 2 saw the Protagonist choose between two objects and repeatedly grasp one; thus, they could attribute a preference to the Protagonist. On the other hand, infants in condition 3 saw the Protagonist repeatedly grasp the only object that was available, and therefore should not necessarily have attributed a preference.
2 Whether or not infants could attribute to the intervening Door-Lifters knowledge of the Protagonist's preference, by varying whether they were onstage during the first phase to observe his grasping actions.

During conditions 1 and 3, the Door-Lifters were onstage to observe the Protagonist's initial grasps; during condition 2 they were not.

Overall, then, condition 1 represented a situation in which there was reason to attribute a preference for the grasped object to the Protagonist, and to attribute knowledge of this preference to the Door-Lifters; it is heretofore referred to as the 'Preference-Knowledge' condition. In condition 2, there was reason to attribute a preference to the Protagonist, but no reason to attribute knowledge of it to the Door-Lifters; it is heretofore referred to as the 'Preference-NoKnowledge' condition. In condition 3, there was no reason to attribute a preference to the Protagonist, and reason to attribute knowledge of this [lack of] preference to the Door-Lifters; it is heretofore referred to as the 'NoPreference-Knowledge' condition. We hypothesized that if our infant observers reason over second-order representational mental states (a fully mentalistic account), they should positively evaluate the Lifter who granted access to the previously grasped toy over the Lifter who granted access to the previously ungrasped toy *only* in the Preference-Knowledge condition, when there was both preference and knowledge information present. If infants instead use a mid-level approach – either assuming that everyone shares the same knowledge of the Protagonist's mental states, or merely representing the first-order goal of the Protagonist and evaluating the Lifter's based on the facilitation/blocking of that goal – they should prefer the Lifter who granted access to the toy the Protagonist had previously grasped in both the Preference-Knowledge and the Preference-Ignorance condition. Finally, if infants are responding only to low-level cues present in the displays, they should fail to distinguish between the Door-Lifters in any condition, as door lifting is present throughout.

We summarize our different models below, grouped into classes 1–4 according to their predictions for the experimental conditions described above. We hereafter refer to the Door-Lifter that granted the Protagonist access to the toy it previously grasped (the preferred toy in conditions 1 and 2; the grasped-but-not-necessarily-preferred toy in condition 3) as 'Grasped-Lifter', and the Lifter who granted the Protagonist access to the toy it did not previously grasp (the dispreferred toy in conditions 1 and 2; the ungrasped-but-not-necessarily-dispreferred toy in condition 3) as 'Ungrasped-Lifter'. The models are as follows:

1 Full Mental model: predicts more infants picking Grasped-Lifter in condition 1 (positively evaluating it as a 'Helper'), but choosing randomly in conditions 2 and 3.
2 Mid-level mental model that assumes second-order goals but not second-order beliefs (referred to as No-Second-Beliefs), or a goal-completion model that takes preference properly into account (Goal-completion1): Both these models predict more infants picking Grasped-Lifter in conditions 1 and 2, but choosing randomly in condition 3.
3 Goal-completion model that does not take preference properly into account (any grasped object is inferred as a goal (Goal-completion2), or feature-based model which uses some feature of the door-lifting behavior to distinguish it as helpful (Feature-based1)). Both these models predict more infants choosing Grasped-Lifter in *all* conditions.
4 Random/Control model, or simple feature-based model that does not distinguish, based on features, between Helpers and Hinderers (Feature-based2): predicts that infants choose randomly in *all* conditions.

We now turn to specifying a computational model that formalizes the Full Mental model. By removing various components, this model can be adapted to capture mid-level models in classes 2 and 3.

Inverse Planning model

Before detailing the experiment and the results, we present a computational model that predicts the inferences an observer reasoning over mental states should make, given the scenarios described above. We explain the model first in intuitive terms, using the experimental set-up as a concrete example, and then give more formal details. The Appendix describes the implementation of the model in depth.

Our model is based directly on previous work examining adults' social evaluations (Ullman *et al.*, 2010, in preparation). It is motivated by the Inverse Planning framework (Baker *et al.*, 2009), which explicitly incorporates the notions of mental states and rational planning. The Inverse Planning framework deals with a general challenge of action understanding: Given observed actions, how can a rational observer (infant or model) reason about the hidden mental states that caused the actions?

To solve this problem, the framework begins by assuming that observers have a representation of other agents as rational planners, i.e. planners that choose actions according to goals and subject to their beliefs about the world. This representation determines the probability distribution for actions, given the hidden goals and beliefs:

$$Rational\ \ Planning \Rightarrow P(A|G,B) \quad (1)$$

where A is the set of possible actions, G the set of possible goals, and B the possible beliefs. For example, consider two particular actions $a1$ and $a2$, where according to the agent's specific beliefs b $a1$ is very likely to lead to the agent's specific goal state g, while $a2$ is unlikely to lead to g. Under a rational-planning probability distribution (Equation 1), $a1$ is relatively likely to be chosen while $a2$ is less likely to be chosen.

This representation can then be used in combination with Bayes' Rule to 'invert' the planning process and reason about the hidden beliefs and goals given a sequence of actions:

$$P(G,B|A)\alpha P(A|G,B)P(G,B) \quad (2)$$

To see how this is useful in a goal-inference case, consider modeling a simple agent as having either goal $g1$ or goal $g2$. By taking action $a1$ the agent will likely achieve goal $g1$, while taking action $a2$ will likely result in achieving goal $g2$. While you as an observer do not have direct access to the agent's goal, you now observe the agent taking action $a1$. By using your model of the agent's planning process, and 'inverting it' through Equation 2, you can reasonably become more certain that the agent has $g1$ as its goal. How certain you ultimately become will depend also on the prior probability you ascribe to the goals, captured in the term $P(G, B)$. One can similarly reason about the beliefs of the agent, or reason jointly about the beliefs and goals of the agent.

This Inverse Planning framework has been used to successfully predict the quantitative and qualitative judgments adults make about agents' goals (Baker *et al.*, 2009) and beliefs and desires (Baker, Saxe & Tenenbaum, 2011), as well as about social goals such as 'helping' or 'hindering' (Ullman *et al.*, 2010, in preparation). An important added assumption in the case of social goals is that helpful agents are defined as agents whose goal it is to bring about the goal states of the agents they are helping, and hindering agents are those whose own goals are to bring about the non-goal states of agents they are hindering. That is, the goal representations of social agents depend on their representations of others' goal representations, rather than being tied to a specific state of the world. Importantly, therefore, social goal representations can be established by a variety of different action types: helping may be cued directly (e.g. by an agent pushing a needy agent toward its goal) or indirectly (e.g. by an agent removing a distal object that blocks the needy agent's goal) (Ullman *et al.*, 2010).

Building on this, we turn to the experimental set-up described in the previous section. As explained earlier, we first need to describe the planning process of the agents, what their goals and beliefs are and where they come from. We model both the Protagonist (the Lion) and the Door-Lifters (Elephants) as agents that take actions until a particular goal state predicate is satisfied. For the Lion, this goal state is either Lion-has-flower or Lion-has-duck. The exact goal of the Lion in each trial is sampled from an underlying preference for either having the flower or the duck, which ranges from 0 to 1. Thus, if the lion has a 0.9 preference for the duck and a 0.1 preference for the flower, it will usually have the duck as its goal in a particular trial and will take the appropriate actions to procure it, but it might sometimes have the flower as its goal.

The *preferences* of the Lion are themselves drawn from a prior distribution, which was chosen to reflect the fact that the observer (or infant) does not know in advance which object the Lion prefers, but it assumes that the Lion has a strong preference for one item over the other. This assumption of an unknown but strong preference can be relaxed without changing the results significantly.

For the Elephant agent, the goal is not a specific world state, but rather it is dependent on the goal of the Lion and inner states of the Elephant, whether it is 'helpful', 'hindering' or 'neutral/random'. A 'helpful' agent thus has as its goal state whatever sets the predicate lion-goal state to true. If in this particular world the Lion wants flowers, a helpful Elephant will help him get a flower. If the Lion instead prefers ducks, the Elephant will help the Lion get a duck. A 'hindering' agent, by contrast, has as its goal state whatever is not the goal state of the Lion. Finally, a 'neutral' or 'random' agent has its own agenda, randomly picking a world state as its goal.

Similar to the Lion's preference, the helping, hindering and neutral qualities of the Elephant form a hidden graded parameter ranging from 0 to 1, and summing up in total to 1. For example, if the helping part of the parameter is set to 1, the goal of the Elephant will always be to satisfy the goal predicate of the Lion, whatever that may be. If it is set to some value between 0 and 1, the Elephant will only sometimes help, to a degree determined by the particular value.

This hidden 'sociability' parameter is drawn from a prior, similar to the Lion's preference. The prior used was uninformative, reflecting the assumption that observers do not know how social the Elephants are ahead of time, nor whether this is a strong tendency.

Having specified the goal predicates for the agents, we define a planning procedure that allows an agent to choose from the actions available to it until a goal predicate is satisfied. The planning procedure was the same for both agents, and basically involves the agent sampling actions conditioned on those actions leading to

its goal state. The exact implementation can be found in the Appendix.

We are now in a position to use Inverse Planning to reason back from observable actions to hidden states like preferences and helpfulness. We start with the simpler case of inferring the preferences of the Lion given the initial phase, and then investigate the model predictions for the actual experimental set-up. In conditions 1 and 2, the first phase has the Lion choosing one object over another. For concreteness, we assume that the Lion reached for the flower. We condition on the Lion reaching for the flower in the presence of the duck from zero to four times, and compute the inferred preference of the Lion. Intuitively, the more times the Lion reaches for the flower over the duck, the more our confidence should grow that it has a strong preference for the flower, and that is indeed the result in Figure 2. For the rest of this section, we assume that the Lion reached four times for the flower, matching what the infants observed (see the Methods section).

We now compare the inferences that a rational observer should make in conditions 1 and 2 in which both the duck and flower are present, with condition 3 in which only one object is present. We again compute the inferred preference of the Lion, keeping the actions the same but varying the world passed into the planning procedure (one with a duck present, the other without a duck present). The results in Figure 3 again agree with intuition: If the Lion chooses the flower when the duck is present this is strong evidence that the Lion has a preference for the flower. If the duck is absent when choosing the flower, no such preference is established.

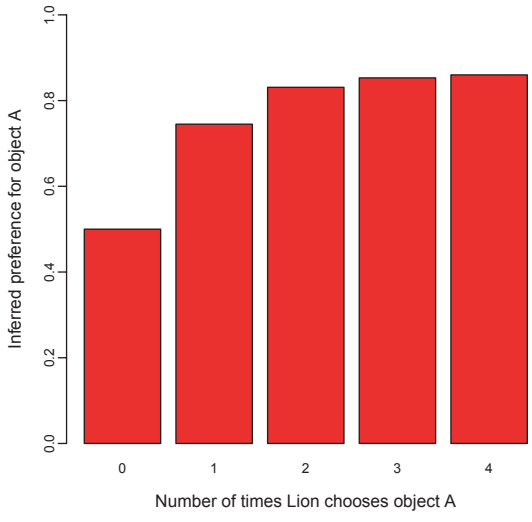

Figure 2 *Inferred lion preference in Preference-Knowledge and Preference-NoKnowledge conditions.*

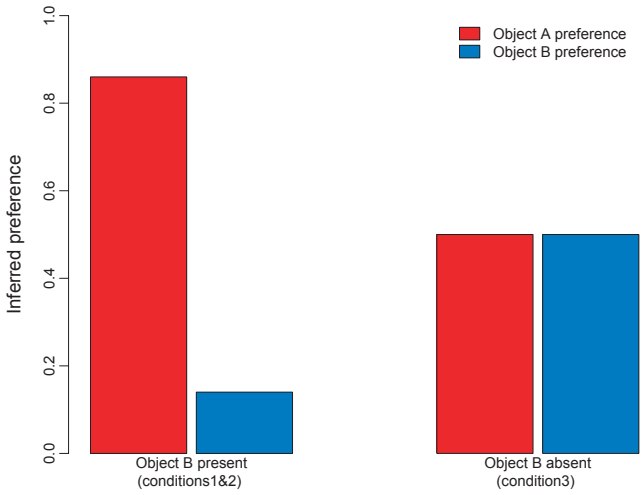

Figure 3 *Inferred lion preference in Preference Knowledge and Preference-NoKnowledge versus NoPreference-Knowledge conditions.*

Having established that the model works reasonably for simple non-social goals, we turn to the experiment. In order to help or hinder, the Elephant must itself infer the preferences of the Lion. We assume that the observer (or infant) models the Elephant as using the same process of computing the Lion's preference that he or she uses to compute this preference. Notice that this is a nested, non-trivial inference. While the observer and the Elephant use the same mental process to infer the preferences of the Lion, it is still possible for the Elephant to reach different conclusions from the observer about this preference, depending on its visual access to the actions of the Lion. If it does have visual access, the Elephant considers the observed actions and infers the preference as described above. If it does not have visual access, the Elephant has no observations and effectively uses the uninformed preference prior, just as an observer who witnessed nothing of the Lion's actions would express ignorance regarding the Lion's preference.

The inference over the hidden states of the Elephant (its knowledge of the Lion's preference, and its social qualities) is similar to the inference over the Lion in terms of the general operations, yet there are important differences and representational demands for the infant in this case. Namely, this inference requires the ability to represent nested inferences (the Elephant reasoning about the Lion), as well as the ability to reason about goals that depend on other goals.

We are specifically interested in the inferred social qualities of the Elephants across the different conditions, asking whether the action of lifting a door will label an Elephant as helping, hindering or neutral. The results of

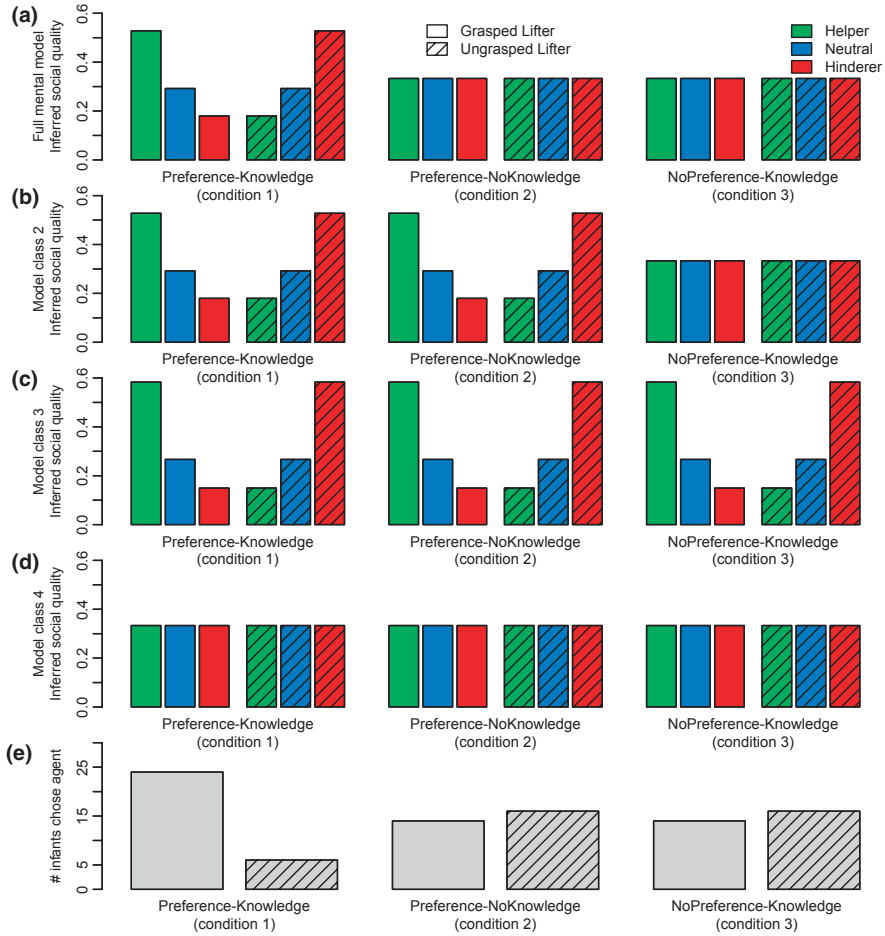

Figure 4 *Model predictions and infant data. Shows the inferred social quality of door-lifters across all conditions, for different models classes. The height of a given bar corresponds to the inferred posterior probability that the elephant will choose that social goal (e.g. to 'Help') in a scenario, given the observed action. (a) Full Mental model; (b) Model class 2, including a second-order goals but not second-order beliefs model, and a goal-completion model that infers preference; (c) Model class 3, including goal-completion model that assumes any grasped object is a goal; (d) control/random model (e) infants' choice of Lifter.*

our model are shown in Figure 4(a). This model makes several important predictions: If the observer can infer the Lion's preference, and the Elephant has visual access which allows it to infer this same knowledge ('Preference-Knowledge' condition), then the Elephant can be evaluated socially depending on the door it opens. If the observer can establish the preference of the Lion, but the Elephant does *not* have visual access allowing it to infer this same knowledge ('Preference-NoKnowledge' condition), then the Elephant cannot be judged as helping or hindering by the door it opens. If the infant and Elephants both have visual access to the Lion's actions, but these are not sufficient to establish a preference ('NoPreference-Knowledge' condition), then the actions of the Elephant are again uninformative. So, the model predicts that the mental states of the Elephant and the Lion (the knowledge of the Elephant and the preference of the Lion) are crucial in socially evaluating perceptually identical actions.

Various parts of this model can be removed to make it capture mid-level mental accounts, instead of the Full-Mental model. We can formalize the predictions of models that do not have second-order beliefs, or those that use goal-completion with preference comprehension (model class 2, as defined previously), by giving the Elephants visual access in all conditions. Doing so equates the belief states of the Elephants with those of the infant in all conditions. Similarly, we can formalize the predictions of a goal-completion model that simply assumes that the grasped object is the goal of the Lion in all cases (part of model class 3), by introducing this exact assumption into the Elephants' goal inferences of the Lion. The inferences of these models are shown in Figure 4(b–c).

We cannot quantitatively assess the predictions of feature-based models in a similar fashion as they do not fall under Inverse Planning, but we can compare them in a quantitative fashion to our model using some simplifying assumptions, discussed in the Results section.

We next present results from an experiment with infants, testing the predictions of this model.

Method

Participants

Ninety full-term 10-month-olds participated (47 girls; mean age = 10 months, 3 days; range = 9;14–10;28). Subjects were randomly assigned to the three conditions, with an approximately equal number of males and females in each group and the same mean age. Thirty-seven additional infants began but were not included in the final sample due to fussiness (21 infants[1]), procedure error (8 infants), and failure to make a choice (8 infants).

Procedure

Preference-Knowledge condition

Infants sat on their parent's lap before a table (W: 122 cm) with a curtain (85 cm from the infants) that could be lowered to occlude the puppet show. A long wall (W: 95 cm; H: 27 cm) constructed of multi-colored blocks spanned the back of the puppet stage. The wall had two openings (W: 16 cm; H: 19 cm); in front of each sat an object (a purple flower (H: 11 cm) and a yellow duck (H: 8 cm)). Two elephant puppets (Door-Lifters; approximately 25 cm high and wearing a yellow and a fuchsia T-shirt) sat at the side of the wall to the infants' left, pointed slightly toward the objects (see Figure 1). Parents were instructed to sit quietly with their infants and not attempt to influence them in any way. Infants were shown 11 events in total.

[1] This is an unusually high fuss out rate for infant puppet show studies (although not for infant studies in general). We hypothesize this was due to the startling nature of the brightly colored wall and the puppet appearing from behind it as if out of nowhere. Indeed, 8 of these fuss outs were during the wall presentation prior to the start of the study or during the first familiarization event; very few fuss outs happened after the familiarization phase was complete. Infants were excluded at an equal rate across conditions (fuss outs: 7/6/8; no choices: 3/3/2; procedure error: 4/2/2 in the Preference-Knowledge, Preference-Ignorance, and NoPreference-Knowledge conditions, respectively).

Toy-grasp Familiarization events

At the start of each trial, the Protagonist (lion) puppet, who had initially been fully occluded by the wall, appeared first behind one opening in the wall, and then the other, apparently 'looking at' the object in front of each. He then appeared back at the first opening, moved through it and grasped the object resting there, and paused. Infants' looking time was recorded online from this point by an observer (who peeked at the infants through a hole in the curtain and could not see the puppet events) until they looked away for 2 consecutive seconds or 30 seconds elapsed, using the program jHab. Between each event, the objects switched locations; the Protagonist performed the same looking and grasping acts (grasping the same object in alternating locations) for four total events.

Door-placement Baseline event

The curtain rose to reveal the elephants resting onstage as in Familiarization; two white doors now blocked the openings in the wall. For half the infants in each condition, the objects were in the location they had been on the last Familiarization event, for the other half the objects switched locations. The Protagonist jumped up and down three times behind the wall such that just the top of his head was visible, to illustrate that he could not get over the wall. After the third jump, action paused (Protagonist invisible) and infants' looking was recorded as above.

Door-Lifting events

The objects were always in the same location as during Baseline and remained there throughout all Door-Lifting events; only one elephant was onstage at a time. Events started with the Protagonist jumping three times as in Baseline. The Door-Lifter then jumped up and moved across the top of the wall to a point above one of the doors, either the door that blocked the previously grasped or the previously ungrasped object. The Lifter then grasped the top of the door, raised it up, and paused, allowing the Protagonist to move through the opening and grasp the object in front. On alternating events, the other elephant opened the other door, and the Protagonist moved through and grasped the other object. Door-Lifting events were alternated for a total of six events; infants' looking was recorded from the Protagonist's grasp as above.

The following were counterbalanced across infants: (1) grasped object (flower/duck); (2) location of grasped object on first Familiarization event (left/right opening);

(3) location of grasped object on Baseline event (same/switched from last Familiarization location); (4) color T-shirt of elephant who lifted the door blocking the previously grasped object (yellow/fuchsia; and (5) order of door lifting (previously grasped object unblocked first/second).

Preference-Ignorance condition

All procedures, counterbalancing, etc., in the Preference-Ignorance condition were as in the Preference-Knowledge condition except that the Door-Lifters were not onstage during Familiarization events. They were onstage during Baseline.

NoPreference-Knowledge condition

All procedures, counterbalancing, etc., in the NoPreference-Knowledge condition were as in the Preference-Knowledge condition except that there was only one object present during the Object-Grasp Familiarization events; thus, when the Protagonist 'looked' through the second opening, there was no object sitting there. The second object was onstage during Baseline.

Choice

Parents turned their chairs 90 degrees to the right so that they were no longer facing the stage, and closed their eyes. An experimenter blind to the identity of the puppets kneeled in front of the infant, said 'Hi [baby's name]' and, while saying 'Look!', held up the puppets, centered on the infant's chest, about 30 cm apart and out of the infant's reach. The infant was required to look at both puppets, and back to the experimenter (prompted by puppet shaking and 'Hi!' if necessary). Once the infant had seen both puppets and the experimenter, the experimenter said 'Who do you like?' and the puppets were moved within reach of the infant. Infants' choices were coded online by this experimenter as the first puppet contacted using a visually guided reach.

The location of the Door-Lifters during choice was counterbalanced across infants within each condition. An independent coder, blind to the identity of the puppets as well as to the experimental condition of the infant, recoded 100% of infants' choices, and agreed with the original experimenter in 100% of cases. In addition, in order to rule out any unconscious bias on the part of the puppeteer (who was necessarily aware of the infants' familiarization condition), a second independent coder, blind to condition, to puppet identity, and to each infant's subsequent puppet choice, watched the puppet shows for each subject's six test trials. She predicted (through a forced-choice) which puppet the puppeteer 'wanted' the baby to choose. Forced choices were unrelated to which puppet was actually helpful in the Preference-Knowledge condition (binomial $p = .86$), and were unrelated to infants' subsequent puppet choices in any condition (binomial ps $> .85$).

Results

In order to assess which of the four alternative models outlined in the Introduction best accounts for infants' social evaluations, we consider a range of different tests with varying degrees of assumptions. We use standard ANOVA and Pearson's chi-square tests to perform a basic analysis of infant's choice and attentional behavior. Next, we use a Bayes factor analysis to compare how well our Full-Mental model is supported by the data, relative to the other models. Finally, we use a permutation test to assess the statistical significance of the Full-Mental model's fit to the data on its own terms.

Choice

All p-values are two-tailed. A Pearson's chi-square test for independence of samples revealed that infants' tendency to choose Grasped-Lifter versus Ungrasped-Lifter differed significantly across the three conditions (χ^2 (2, N = 90) = 9.11, $p = .01$, $\omega = .46$). Infants in the Preference-Knowledge condition significantly preferred Grasped- to Ungrasped-Lifter (24 out of 30 infants; binomial $p = .001$), whereas infants chose equally between Lifters in both the Preference-Ignorance condition (14 of 30 chose the Grasped-Lifter; binomial $p = .86$) and the NoPreference-Knowledge condition (14 of 30 chose Grasped-Lifter; binomial $p = .86$).

There were no effects of sex, preference object, side of preference object during the first familiarization trial, puppet shirt color, location of Lifters during familiarization, whether or not the location of the grasped toy switched sides during the baseline event, the order of Grasped- and Ungrasped-Lifting events, or whether the Grapsed-Lifter lifted to door nearer to it (on the infants' left) or farther from it (on the infants' right) on choice within or across conditions (all Fisher's Exact ps $> .05$).

Attention to puppet events

Infants' attention to all events in each condition is depicted in Table 1. Summaries for different event-types are detailed below.

Table 1 *Infants' attention to each event in seconds (SEM); both within and across conditions*

	Preference-Knowledge	Preference-Ignorance	NoPreference-Knowledge	Totals
Toy Grasp 1	15.62 (1.51)	13.63 (1.20)	13.39 (1.20)	14.21 (.75)
Toy Grasp 2	11.77 (1.56)	9.83 (1.33)	11.42 (1.26)	11.06 (.78)
Toy Grasp 3	9.80 (1.28)	8.60 (1.56)	7.27 (.79)	8.62 (.70)
Toy Grasp 4	8.28 (1.20)	9.55 (1.40)	9.46 (1.23)	9.07 (.71)
Baseline (BL)	10.83 (1.25)	12.32 (2.31)	13.01 (1.29)	12.11 (.76)
Switch BL	9.33 (1.58)	15.96 (1.27)	14.25 (2.11)	13.18 (1.22)
Stay BL	12.33 (1.90)	9.05 (1.19)	11.77 (1.49)	11.05 (.90)
Grasped-Lifter 1	11.08 (1.29)	7.43 (.46)	8.82 (1.14)	9.10 (.70)
Ungrasped-Lifter 1	10.2 (1.35)	9.38 (1.01)	7.8 (1.04)	9.23 (.70)
Grasped-Lifter 2	7.07 (.95)	6.07 (.73)	5.92 (.70)	6.41 (.46)
Ungrasped-Lifter 2	8.05 (1.29)	6.91 (.79)	6.13 (.66)	7.08 (.56)
Grasped-Lifter 3	6.26 (.9)	7.07 (1.24)	5.92 (.89)	6.45 (.57)
Ungrasped-Lifter 3	7.30 (1.18)	6.10 (1.01)	5.69 (.77)	6.33 (.57)

Toy-grasp Familiarization events

Infants' attention to Familiarization events did not differ by condition, either by trial (multivariate ANOVA on attention to each of the four familiarization events, all $F(2, 87) < 1.2$; all $ps > .33$; all $\eta_p^2 < .03$), or summed across trials (univariate ANOVA on attention to the summed looking on all 4 familiarization events, $F(2, 76) = .42$, $p = .66$, $\eta_p^2 = .01$). Across conditions, infants looked an average of 14.21 seconds to the first toy-grasp, and 9.07 seconds to the fourth (last) toy-grasp, reflecting a significant decrease in looking time over familiarization (paired $t(98) = 5.71$, $p = .000$, $\eta^2 = .27$).

Door-placement Baseline event

Attention to the Baseline event did not differ by condition ($mean_{Baseline} = 12.11$ s (SEM = .76); $F(2, 87) = .74$ $p = .48$, $\eta_p^2 = .02$). Infants looked longer to Baseline events than to the fourth toy-grasp event (paired $t(89) = -3.76$; $p = .000$, $\eta^2 = .14$), suggesting they noticed the change.

Switch versus Stay Baseline events

Although infants looked approximately 2 seconds longer to baseline events in which the toys switched places (mean = 13.18 s) than those in which the toys stayed put (mean = 11.05 s), this did not reach significance ($F(1, 88) = 1.96$; $p = .17$; $\eta_p^2 = .02$).

Door-Lifting events

Infants across conditions looked equally over the three Grasped-Lifter and three Ungrasped-Lifter events ($mean_{Grasped} = 21.95$ s (SEM = 1.19); $mean_{Ungrasped} = 22.63$ s (SEM = 1.36); paired $t(89) = -.54$, $p = .59$). These measures did not differ by condition (repeated-measures ANOVA, $F(2, 87) = .49$, $p = .62$; $\eta_p^2 = .01$).

Attention's effect on Choice

A univariate ANOVA on infants' choice of Grasped-Lifter (assigned a value of 1) versus Ungrasped-Lifter (assigned a value of 0) with total time to toy-grasping Familiarization, Baseline, Grasped-Lifter, and Ungrasped-Lifter events as covariates revealed no significant effects across or within condition (all $ps > .05$).

Model comparison

We use a Bayes factor (BF) analysis to contrast between the four model classes presented in the introduction. A Bayes factor analysis computes the marginal likelihood of the data under two different models (e.g. model 1 and model 2), and compares them using the ratio:

$$K = \frac{P(data|model_1)}{P(data|model_2)}$$

The data provide clearer support for model 1 or 2 to the extent that K is greater than or less than 1, respectively.

In order to compute these marginal likelihoods, we must move from the qualitative predictions of the different model classes to quantitative predictions about the number of infants expected to choose either agent (recognizing that even if a model is 'correct', none of these models is complete and hence not all participants should be expected to follow its predictions in any given situation). We use a general approach to compare the predictive value of different models, by considering infants' choices of Grasped-Lifter or Ungrasped-Lifter agents as being drawn from a coin-flip with a weight θ that embodies each model's predictions. This coin weight represents the expected proportion of infants choosing

the Grasped-Lifter agent in a given condition, with different models assigning different values to this weight in different conditions. The likelihood of a sequence of choices in which k out of N infants choose the Grasped-Lifter is:

$$\theta^{\kappa}(1-\theta)^{\mathrm{N}-\kappa}$$

The Full-Mental and several mid-level models provide a specific numerical result for the inference of social traits (as shown in Figure 4), but this needs to be translated into a coin weight θ. It seems intuitive to use the ratio between the inferred 'helpfulness' or 'hindering-ness' of each agent as the predicted coin weight. For the Full-Mental model this gives an expected proportion of about 75% of infants choosing the Grasped-Lifter agent in condition 1, and 50% for the other conditions (one can use either the helpfulness ratio or the hindering-ness ratio as they are symmetric here). For models in class 2, we have an expected proportion of about 75% for conditions 1 and 2, and 50% for condition 3. For the goal-completion model in class 3, we have an expected proportion of about 80% for all conditions. Model class 4 (the 'random/control' class) predicts an expected proportion of 50% in all cases.

Using the data from the different conditions, we can now compute the Bayes factor K comparing all pairs of models described above. We find:

1. Model 1 vs. Model class 2 (Full-Mental vs. No-Second-Beliefs/Goal-Completion1):

   K = 187:1

2. Model 1 vs. Model class 3 (Full-Mental vs Goal-Completion2):

   K = 246599:1

3. Model 1 vs. Model class 4 (Full-Mental vs. Control/Random):

   K = 254:1

Thus, a Bayes factor analysis finds the evidence strongly in favor of our model. However, we did not consider various feature-based models, as it is not clear what their precise prediction for the coin weight θ would be. One could also argue the various model classes represent whole classes of approaches that could be implemented in other ways.[2] The most general way that we could relate each of the four models' qualitative predictions to exact quantitative predictions of the coin bias θ is to specify that in some cases θ should be 1/2 and in others it should be greater than 1/2.

Therefore, in order to provide a fair comparison between all models, we do the following: for all conditions in which a model class predicts more infants choosing Grasped-Lifter, we use a marginal likelihood that integrates out over possible values of the coin weight θ in the range [1/2, 1], signifying that the proportion of infants choosing the Grasped-Lifter is expected to be above chance, but without committing to the strength of the effect. In conditions in which a model group predicts random choice, we use $\theta = 1/2$.

We again compute the Bayes factor K comparing all pairs of models, and find:

1. Model 1 vs. Model class 2 (Full Mental vs. No-Second-Beliefs/Goal-Completion1):

   K = 5.8:1

2. Model 1 vs. Model class 3 (Full-Mental vs. Goal-Completion2/Feature-Based1):

   K = 33.9:1

3. Model 1 vs. Model class 4 (Full-Mental vs. Control/Random/Feature-Based2):

   K = 116.6:1

Thus, a Bayes factor model analysis finds the evidence in favor of the Full-Mental model, in comparison to the other models considered, including the feature-based models and with few assumptions about the strength of the effect. It strongly rejects model classes 3 and 4, and supports our model in comparison to model class 2.

Permutation test

A permutation test allows us to test the fit of the Full-Mental model while making few assumptions about the distribution being used by infants to pick Lifters. We first define a test statistic measuring fit under a given model. We then consider all possible rearrangements of the data – that is, we take the total number of infants who chose Grasped-Lifter, and reassign them to the three conditions in all possible permutations. We recomputed the value of the statistic for each such permutation. The proportion of times that the value of the statistic is greater than that obtained under our model, gives us a one-sided p-value. This test only assumes that it is possible the infants in all groups are not affected by the experimental manipulations, and makes no assumptions about the type of distribution by which infants' choices are made (unlike a chi-square test, for example).

[2] It is also possible to argue that an infant can use the combined probability of an agent being a helper+neutral, or a hinderer+neutral to make a choice. These give an expected proportion between 65% and 80%, and do not change the analysis substantially.

The Full-Mental model predicts that children will be at chance in conditions 2 and 3, but will choose Grasped-Lifter more in condition 1. Therefore, if we designate choosing Grasped-Lifter as '1' and choosing Ungrasped-Lifter as '0', we can consider the following test statistic:

$$|mean(condition_1) - 38; 0.5| - |mean(condition_2) \\ 38; -0.5| - |mean(condition_3) - 0.5|$$

Notice that the more the proportion of infants who chose the Grasped-Lifter in conditions 2 and 3 deviates from chance, or the more the this value approaches chance in condition 1, the smaller the statistic. Thus, by checking the number of permutations in which the value of the statistic in greater than that obtained in our data, we can assess whether our particular data-set happened by chance, or whether the infants are sensitive to the experimental manipulations as predicted.

The total number of possible permutations is too large to enumerate exhaustively, but using a large sample of 10^6 random permutations, we find that $p < .05$ (specifically, $p = .018$).

In sum, all of the analysis methods used – standard binomial tests and a Pearson's chi-square, Bayes factor model comparisons, and a permutation test for goodness of model fit – converge on the hypothesis that the Full-Mental model accounts best for infants' choices.

Discussion

Infants' choices across conditions were exactly as the mentalistic model predicts; indeed, the Full-Mental model explains the data better than three other groups of models involving varying degrees of mental states. When the Protagonist expressed a preference, and when the Door-Lifters were present to observe that preference, infants preferred the Door-Lifter who gave the Protagonist access to its preferred toy. This result is reminiscent of previous research in which infants prefer those who help versus hinder third parties in their goals (e.g. Hamlin & Wynn, 2011; Hamlin *et al.*, 2007, 2010), and suggests that they positively evaluated the helpful Lifter and/or negatively evaluated the unhelpful one. In contrast, both when the Protagonist expressed a toy preference but the Door-Lifters were absent and therefore lacked knowledge of that preference, and when the Door-Lifters were present to observe that the Protagonist did not express a preference, infants chose randomly between the Door-Lifters, suggesting that they did not differentially evaluate them.

These results argue strongly against the idea that infants rely solely on perceptual cues when evaluating social actions between agents. That infants failed to distinguish the characters in the NoPreference-Knowledge condition suggests that they were truly responding to the *goal* of the Protagonist: if infants respond merely to some physical cues present within pro- and antisocial behaviors – for example, facilitating a former action sequence – they should have preferred the Lifter who opened the door blocking the previously grasped toy in all conditions, as the Grasped-Lifter always facilitates the Protagonist's former action sequence. If, instead, a low-level cue was present within Door-Lifting itself, infants should not have distinguished the Lifters in any condition, as lifting was virtually identical across Lifters (and lifting a near versus far door did not influence infant's choices within any condition).

Similarly, these results also argue forcefully against both mid-level accounts introduced. If infants respond only to a Protagonist's *first-order goals*, and positively evaluate any individual who completes them, they should have chosen the Lifter who gave access to the grasped toy in both conditions in which a goal (in this case, a preference) was inferable: the Preference-Knowledge condition and the Preference-Ignorance condition. Similarly, if infants can infer that the Door-Lifters have second-order goals to help/harm, but do not also represent that intentionally helping/harming someone requires having *knowledge* of what his goal is, they should have chosen the Lifter who gave access to the preferred toy in both Preference conditions.

In sum, we reject both a cue-based and several mid-level accounts of infants' preferences for the Grasped-Lifter in the current experiments, and believe that these results suggest that infants are capable of mentalistic social evaluations in general, in which helping is represented as the intention to aid others' goals, and/or harming is represented as the intention to prevent others' goals. Of course, there remain some unanswered questions. First, it is unclear that infants were *both* representing helping as an intention to aid and harming as an intention to prevent. Infants' preference for the Grasped-Lifter to the Ungrasped-Lifter in the Preference-Knowledge condition may have been due to a preference for the helpful character, an aversion to the unhelpful character, or both. While previous results suggest that infants *both* prefer helpful to neutral and neutral to unhelpful individuals at this age (Hamlin *et al.*, 2007), we did not test this in the current studies. The mentalistic model assumed a distinction natural to adults, between helping, hindering and neutral intentions, and can theoretically distinguish between all of them given a wider range of actions on the social agents' part. An addition unexplored question is whether infants could incorporate others' *beliefs* into their evaluation of helping and

hindering. In addition to the presence or absence of knowledge of others' goals, an understanding of what others believe they are doing is often incorporated into adult's evaluations. For example, one sometimes *thinks* they are helping, but actually behaves neutrally or negatively; it is up to future studies to pursue these vital aspects of adult social evaluation.

The current experiments combined a study of infants with a modeling approach; we found that the model predicts our data and believe that each method complements the other. While the infant studies presented here are clearly the result of what we hypothesize to be driving early social evaluation, and a large body of research with adults confirms that analyses of mental states drives moral evaluation in adulthood (e.g. Guglielmo, Monroe & Malle, 2009), to assume that infants' evaluations are also driven by complex mentalistic analyses would be premature. Indeed, a large body of research on the development of moral reasoning suggests that even children many years older than our infant subjects sometimes have difficulty incorporating mental state analyses into their judgments (e.g. Piaget, 1932/1965; see Karniol, 1978, for a review) and a host of results show a developmental trend toward increasing effects of mental state analysis with age (e.g. Buchanon & Thompson, 1973; Costanzo, Coie, Grumet & Farnill, 1973; Gutkin, 1972; Hebble, 1971; Helwig, Hildebrandt & Turiel, 1995; Leon, 1980; Surber, 1977; Zelazo, Helwig & Lau, 1996). Thus, it is especially important to formally model exactly what is driving judgments in this task, and to determine whether the model fits the observed infant data. We add this to a trend of using an Inverse Planning approach with an intuitive psychology viewpoint to describe infants' social understandings specifically (e.g. Baker, Tenenbaum & Saxe, 2009; Ullman *et al*., 2010, in preparation) as well as combining probabilistic modeling with cognitive development research in general (e.g. Perfors, Tenenbaum, Griffiths & Xu, in press; Téglàs, Vul, Girotto, Gonzalez, Tenenbaum & Bonatti, 2011; Xu, 2007; Xu & Griffiths, in press; Xu & Tenenbaum, 2007; Gopnik, Wellman, Gelman & Meltzoff, 2010).

Acknowledgements

We thank the parents and infants who participated, and the members of the Centre for Infant Cognition at the University of British Columbia. This work was supported by a Natural Sciences and Engineering Research Council of Canada grant to the first author, and an Office of Naval Research (ONR) grant to the third and fourth authors.

References

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

Baker, C.L., Saxe, R.R., & Tenenbaum, J.B. (2011). Bayesian theory of mind: Modeling joint belief–desire attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Baker, C.L., Goodman, N.D., & Tenenbaum, J.B. (2008). Theory-based social goal inference. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, 1447–1452.

Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science*, **8**, 492–499.

Biro, S., & Leslie, A.M. (2007). Infants' perception of goal-directed actions: development through cue-based bootstrapping. *Developmental Science*, **10**, 379–398.

Blythe, P.W., Todd, P.M., & Miller, G.F. (1999). How motion reveals intention: categorizing social interactions. In G. Gigerenzer & P. Todd. (Eds.), *Simple heuristics that make us smart* (pp. 257–285). Oxford: Oxford University Press.

Brandone, A., & Wellman, H. (2009). You can't always get what you want: infants understand failed goal-directed actions. *Psychological Science*, **20** (1), 85–91.

Buchanon, J.P., & Thompson, S.K. (1973). A quantitative methodology to examine the development of moral judgment. *Child Development*, **44**, 186–189.

Costanzo, P.R., Coie, J.D., Crumet, J.F., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development*, **44**, 154–171.

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

Dunfield, K.-A., & Kuhlmeier, V.-A. (2010). Intention-mediated selective helping in infancy. *Psychological Science*, **21**, 523–527.

Gao, T., Newman, G.E., & Scholl, B.J. (2009). The psychophysics of chasing: a case study in the perception of animacy. *Cognitive Psychology*, **59** (2), 154–179.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: infants' reactions to equal and unequal distributions of resources. *Developmental Science*, **14**, 1012–1020.

Gopnik, A., Wellman, H., Gelman, S., & Meltzoff, A. (2010). A computational foundation for cognitive development: comment on Griffiths et al. and McLelland et al. *Trends in Cognitive Sciences*, **14** (8), 342–343.

Guglielmo, S., Monroe, A.E., & Malle, B.F. (2009). At the heart of morality lies folk psychology. *Inquiry*, **52**, 449–466.

Gutkin, D.C. (1972). The effect of systematic story changes on intentionality in Children's moral judgments. *Child Development*, **43**, 187–195.

Hamlin, J.K., Newman, G., & Wynn, K. (2009). Eight-month-old infants infer unfulfilled goals, despite ambiguous physical evidence. *Infancy*, **14** (5), 579–590.

Hamlin, J.K., & Wynn, K. (2011). Five- and 9-month-old infants prefer prosocial to antisocial others. *Cognitive Development*, **26**, 30–39.

Hamlin, J.K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, **450**, 557–559.
Hamlin, J.K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, **13**, 923–929.
Hebble, P.W. (1971). Development of elementary school children's judgment of intent. *Child Development*, **42**, 583–588.
Helwig, C.C, Hildebrandt, C., & Turiel, E. (1995). Children's judgments about psychological harm in social context. *Child Development*, **66**, 1680–1693.
Karniol, R. (1978). Children's use of intention cues in evaluating behavior. *Psychological Bulletin*, **85**, 76–85.
Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, **14**, 402–408.
Kushnir, T., Xu, F., & Wellman, H.M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science*, **21**, 1134–1140.
Leon, M. (1980). Coordination of intent and consequence information in children's moral judgment. In F. Wilkening, J. Becker & T. Trabasso (Eds.), *Information integration by children* (pp. 71–112). Hillsdale, NJ: Erlbaum.
Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent. *Developmental Science*, **14**, 453–460.
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science*, **16**, 601–608.
Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, **105**, 489–512.
Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, **12**, 142–149.
Marsh, H.L., Stavropoulos, J., Nienhuis, T., & Legerstee, M. (2010). Six- and 9-month-old infants discriminate between goals despite similar action patterns. *Infancy*, **15**, 94–106.
Onishi, K., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, **308** (5719), 255–258.
Perfors, A., Tenenbaum, J.B., Griffiths, T.L., & Xu, F. (in press). A tutorial introduction to Bayesian models of cognitive development. *Cognition*.
Piaget, J. (1932/1965). *The moral judgment of the child*. New York: Free Press.
Premack, D., & Premack, A.J. (1997). Infants attribute value ± to the goal-directed actions of self-propelled objects. *Journal of Cognitive Neuroscience*, **9**, 848–856.
Ruffman, T., & Perner, J. (2005). Do infants really understand false belief? Response to Leslie. *Trends in Cognitive Sciences*, **9**, 462–463.
Scholl, B.J., & Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, **4** (8), 299–309.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, **18** (7), 587–592.
Surber, C.F. (1977). Developmental processes in social inference: averaging intentions and consequences in moral judgment. *Developmental Psychology*, **13**, 654–665.
Surian, L., Caldi, J., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, **18** (7), 580–586.
Tésd, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J., & Bonatti, L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, **332** (6033), 1054–1059.
Tomasello, M., & Haberl, K. (2003). Understanding attention: 12- and 18-month-olds know what is new for other persons. *Developmental Psychology*, **39** (5), 906–912.
Ullman, T.D., Baker, C.L., Macindoe, O., Evans, O., Goodman, N.D., & Tenenbaum, J.B. (2010). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, **22**, 1874–1882.
Ullman, T.D., Macindoe, O., Baker, C.L., Evans, O., Goodman, N.D., & Tenenbaum, J.B. (in preparation). Bayesian models of social goal inference.
Woodward, A. (1998). Infants selectively encode the goal of an actor's reach. *Cognition*, **69**, 1–34.
Woodward, A.L. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, **22** (2), 145–160.
Xu, F. (2007) Rational statistical inference and cognitive development. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind: Foundations and the future* (Vol. **3**, pp. 199–215). Oxford: Oxford University Press.
Xu, F., & Griffiths, T. (in press). Probabilistic models of cognitive development: towards a rational constructivist approach to the study of learning and development. *Cognition*.
Xu, F., & Tenenbaum, J.B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, **10**, 288–297.
Zelazo, P.D., Helwig, C.C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, **67**, 2478–2492.

Received: 10 November 2011
Accepted: 11 July 2012

Appendix A

Specification of the model

As mentioned in the modeling section, the overall framework used is that of Inverse Planning. We assume that observers represent other agents as rational planners, having some way of choosing actions according to their goals and based on their beliefs. This representation gives a probability distribution for choosing an action at a given state, given the agent's beliefs and goals:

$$Rational\ \ Planning \Rightarrow P(A|S, G, B) \qquad (1)$$

where A is the set of possible actions, G the set of possible goals, and B the possible beliefs. Such a

representation can then be inverted in a Bayesian fashion, going from the visible action and state sequence to reason about hidden variables such as goals and beliefs:

$$P(G, B|A, S)\alpha P(A|S, G, B)P(G, B) \quad (2)$$

There are many possible ways to implement rational planning and run inverse inference. In this paper, we implemented an 'action-as-conditioned-sampling' procedure, and use the Church (Goodman, Mansighka, Roy, Bonawitz, & Tenenbaum, 2008) probabilistic programming framework to do inference. We first describe the planning procedure, then the generative procedure used by the observer in general. We then turn to inference.

The planning procedure takes in an agent and a transition function. It samples an action from the agent prior over actions, conditioned that action leading to the desired goal-state. The procedure then returns the sampled action. More concretely, if an agent has a prior distribution over actions $P(A = a_i)$, this procedure gives the distribution:

$$\begin{aligned} &P38;38;(A = a_i|a_i \text{ leads to state } S_{goal}) \\ &38;38;= \frac{T(S_{goal}|S, A = a_i)P(A = a_i)}{\sum_j T(S_{goal}|S, A = a_j)P(A = a_j)} \end{aligned} \quad (3)$$

Where T is the transition function that returns a distribution over next states, given the current state and action. The procedure can be easily adapted for chains of actions, although it makes equation (3) more involved.

The goal of an agent on a given trial, in a specific environment, was itself drawn from the preferences of the agent, represented by a multinomial distribution:

$$P(goal_i) \sim Multinomial(\theta_i) \quad (4)$$

Thus, if for example θ_i is high then $goal_i$ is likely to be the goal of the agent in a given situation. For the simple agent, the goals were functions of different states of the world, e.g. $goal_1$(state) returned 'True' if state = Lion_has_flower. For the social agent, the goal was a function of the state and the goal of the other agent, e.g. $goal_{helpful}$ = goal(Lion) is a function which takes in a state and returns 'True' if that state is the goal of the Lion.

These θ parameters of the multinomial distribution were drawn from a prior discretized Dirichlet distribution, and they are the main target of inference. For the Lion agent, they represent a preference for flowers or ducks. For the Elephant agent, they represent the tendency to choose the goal function 'Help', 'Hinder' or 'Neutral', which we refer to as the 'social qualities' of the Elephant.

In the case of the Lion, because there were only two possible goals, the Dirichlet distribution turned into a discretized beta, chosen with parameters to reflect the fact that the Lion likely has a high preference for one of the two objects, but it is unclear which. The discretization was such that the multinomial parameters were either ($\theta_{flower} = 0.1$, $\theta_{duck} = 0.9$), ($\theta_{flower} = 0.9$, $\theta_{duck} = 0.1$), ($\theta_{flower} = 0.2$, $\theta_{duck} = 0.8$), or ($\theta_{flower} = 0.2$, $\theta_{duck} = 0.8$). The discretized beta distribution can be made more uniform without altering the results significantly. In the case of the Elephant, the discretized Dirichlet was uninformative, representing the fact that we are unsure what the social tendencies of the Elephant are a-priori.

Combining everything above, the generative part of the way the observer represents the Lion agent is:

1. Draw the parameters θ of a binomial preference distribution from the discretized beta.
2. Draw a specific goal from the preference distribution.
3. Use the planning procedure to draw an action for the Lion, given the goal and transition function.

The generative part of the way the observer represents the Elephant agent is:

1. Infer the preference of the Lion (depending on visual access).
2. Draw a specific goal for the Lion from the inferred preference.
3. Draw a multinomial distribution over social goals for the Elephant from the discretized Dirichlet.
4. Draw a specific social goal from the social goal distribution.
5. Use the planning procedure to draw an action for the Elephant, given the social goal and transition function.

These generative procedures lead to a distribution on the variables that appear in them, including the parameters of interest θ. To run inference, we condition on the action variables receiving certain assignments, thus taking into account observations and changing the distribution to a posterior over the variables. This inference can be done by direct sampling, but we preferred to use Cosh – a dynamic programming implementation of the Church probabilistic programming language. Given a Church program, Cosh computes the program's distribution on return values (the marginal distribution) exactly. Cosh turns the program into a system of polynomial equations that has the marginal probabilities as solutions, and then solves this system to obtain the marginal distribution. The result of this inference for various conditions is presented in Figures 2–4, as described by the main text.

Appendix B

Specifics of the model as a probabilistic program

```
;;;CONSTANTS
  ; probability of choosing an action regardless of goal
  (define no-goal-epsilon 0.000001)
  ;; ***** End of constants *****
  ;;; AGENT CONSTRUCTION AND SELECTORS
  ;; Self-explanatory in terms of the construction.
  ;; this makes it easier to pass around agents
  ;; especially in a multi-agent setting.
  (define (make-agent name goal action-prior)
  (list name goal action-prior))
  (define (get-agent-name agent) (first agent))
  (define (get-agent-goal agent) (second agent))
  (define (get-agent-action-prior agent) (third agent))
  ;; ***** End of agent construction *********
  ;;; LION AGENT CONSTRUCTION
  ;; The lion as a prior chooses uniformly between
  ;; the actions available to it:
  ;; In this experiment, these are getting the duck, getting the flower.
  (define lion-possible-actions '(lion-get-duck lion-get-flower))
  (define (lion-action-prior) (uniform-draw lion-possible-actions))
  ;; NOTE 1: the following is NOT the goal, this
  ;; is the construction of the goal predicate.
  ;; For example, if the goal prior is (0.9 flower, 0.1 duck), then the
  ;; lion might end up with the goal of either flower or duck, it just
  ;; has a higher probability of going for the flower. This goal-prior
  ;; is essentially the long-term preference.
  ;; NOTE 2: The predicate is drawn anew each time
  (define (choose-lion-goal preferences)
  (lambda (world-state)
  (let ((goal-food (sample preferences)))
  (equal? world-state goal-food))))
  ;; ***** End of LION function definitions **********;;
  ;;;ELEPHANT AGENT CONSTRUCTION
  ;; The ELEPHANT as a prior chooses uniformly between the actions available to it:
  ;; In this experiment, these are opening and closing gates.
  (define elephant-possible-actions '(elephant-open-right elephant-open-left))
  (define (elephant-action-prior) (uniform-draw elephant-possible-actions))
  ;;The elephant either adopts the lion's goal, or a predicate of NOT(lion's goal)
  ;;helping-hindering is also decided each time anew, drawn from prior.
  ;;So, for example, if the helper-prior is (0.9 help, 0.1 hinder) the elephant
  ;;might still hinder, it's just less likely. The prior is the quality we will
  ;; be querying over as the elephant's general 'helpfulness'.
  (define (elephant-goal-predicate lion-goal-predicate helpfulness)
  (lambda (state)
  (let ((elephant-disposition (sample helpfulness)))
  (if (equal? elephant-disposition 'helper)
  (lion-goal state)
  (if
  (equal? elephant-disposition 'hinderer)
  (not (lion-goal state))
  #t))))
  ;; ***** End of ELEPHANT function definitions
  ;;; DIFFERENT WORLD TRANSITIONS
  ;; Condition 1&2: Both objects are in play and retrievable.
  (define (world-condition1&2-transition action)
  (case action
  ((lion-get-duck) 'lion-has-duck)
  ((lion-get-flower) 'lion-has-flower)
  ; Any other action stays in opening position
  (else 'start)))
  ;; Condition 3: Only one object is in play and retrievable
  (define (world-condition3-transition action)
  (case action
  ((lion-get-flower) 'lion-has-flower)
  ; Any other action stays in opening position
  (else 'start)))
  ;; Setup of world A.
  ;; Both objects are in play:
  ;; duck behind left, flower behind right.
  ;; actions for the elephant: one for opening each gate.
  ;; The experiment is such that the lion retrieves whatever
  ;; is behind the gate, so there's no added rollout.
  (define (world-A-transition action)
  (case action
  ((elephant-open-right) 'lion-has-flower)
  ((elephant-open-left) 'lion-has-duck)
  (else 'gates-closed)))
  ;; ******* End of different world transitions ********;;
```

```
;;; ACTION PLANNING AS QUERY
;; single agent planning
(define (planning-procedure world agent)
(rejection-query
; get the goal
(define goal? (get-agent-goal agent))
; take an action
(define action ((get-agent-action-prior agent)))
; find the resulting state
(define end-state (world action))
;; query on the action
action
;; condition on getting to the goal
(or (goal? end-state) (flip no-goal-epsilon))))
;; ****** End of action planning as query **********
;;; GOAL INFERENCE AS QUERY
;; single agent goal/reward inference
;; The inference of the lion's goal
(define (infer-goal-lion observed-action world-trans)
(rejection-query
;; prior over goals is drawn from a discretized beta showing preference for some item
;; but unclear which item exactly
(define preference-prior (uniform-draw '(0.1 0.2 0.8 0.9)))
(define (food-prior) (multinomial '(lion-has-flower lion-has-duck)
(list preference-prior (- 1 preference-prior))))
;; after the goals are defined we construct the appropriate goal functions and agent
(define lion-goal (make-lion-goal food-prior))
(define lion (make-agent 'lion' lion-goal lion-action-prior))
;; sample a state-movement sequence for the lion given the goal functions
(define (sampled-action)
(sample-state-action-single-agent world-trans lion))
;; query on the goal preferences
(food-prior)
;; conditioned on the action being equal to the observed one
;; multiple reaches can be simulated.
(and (equal? observed-action (sampled-action))
(equal? observed-action (sampled-action))
(equal? observed-action (sampled-action))
(equal? observed-action (sampled-action)))
))
;; use this in to model Goal completion, model class 3
(define (model-class-3-infer-food-prior observed-lion-action)
(if (equal? observed-lion-action 'lion-get-flower)
'(1.0 0.0)
'(0.0 1.0)))
;; multi-agent goal inference
;; The inference of the elephant's goal/disposition
(define (infer-helpfulness-elephant observed-elephant-action observed-lion-action
world-elephant-lion world-lion-alone
elephant-had-visual-access?)
(rejection-query
; Comment this in to model class 3 models, strict goal completion
;(define (inferred-food-prior)
;(multinomial '(lion-has-flower lion-has-duck) (model-class-3-infer-;food-prior observed-lion-action)))
; If the elephant had no access, assume uniform prior
(define (inferred-food-prior)
(if elephant-had-visual-access?
(infer-goal-lion observed-lion-action world-lion-alone)
(multinomial '(lion-has-flower lion-has-duck) '(0.5 0.5))))
;; after the goal prior is defined we construct the appropriate goal functions and agent
(define inferred-lion-goal (make-lion-goal inferred-food-prior))
(define inferred-lion (make-agent 'lion' inferred-lion-goal lion-action-prior))
;; define priors on elephant's helpfulness, the elephant goal and the agent itself
(define helper-weight (discrete-dirichlet-3))
(define (helper-prior) (multinomial '(helper hinderer neutral) helper-weight))
(define elephant-goal (make-elephant-goal inferred-lion-goal helper-prior))
(define elephant (make-agent 'elephant' elephant-goal elephant-action-prior))
;; sample a state-movement sequence for the elephant
(define (sampled-elephant-action)
(sample-state-action-single-agent world-elephant-lion elephant))
;; query on the helpfulness
(helper-prior)
;; conditioned on the action being equal to the observed one
(and (equal? observed-elephant-action (sampled-elephant-action))
(equal? observed-elephant-action (sampled-elephant-action))
(equal? observed-elephant-action (sampled-elephant-action)))
))
;; ******* End of goal inference as query **********
;;; THE EXPERIMENT
;;; Observations
```

```
;; observed actions establishing preference
(define observed-lion-action 'lion-get-flower)
(define observed-elephant-action 'elephant-open-right)
;;; Conditions
;; Condition 1: The baby and elephant observe the lion choosing a flower over a duck
;; Condition 2: The baby observes the lion choosing a flower over a duck, but the
;;
elephant doesn't
;; Condition 3: The baby and elephant observe the lion choosing a flower, no duck present
(define condition 'condition3)
;; COMMENT IN THE FOLLOWING TO CAPTURE A CLASS 2 MODEL
;;(define elephant-had-visual-access? #t)
(define elephant-had-visual-access?
(if (or (equal? condition 'condition1) (equal? condition 'condition3))
#t
#f))

(define lion-alone-world-transition
(if (or (equal? condition 'condition1) (equal? condition 'condition2))
world-condition1&2-transition
world-condition3-transition))
(define lion-and-elephant-world-transition world-A-transition)
;;; Queries
;; Query 1: What is the lion's preference?
;(infer-goal-lion observed-lion-action lion-alone-world-transition)
;; Query 2: What is the elephant's helpfulness?
(infer-helpfulness-elephant observed-elephant-action observed-lion-action
lion-and-elephant-world-transition lion-alone-world-transition
elephant-had-visual-access?)
```

Reference: N. D. Goodman, V. K. Mansighka, D. Roy, K. Bonawitz, J. B. Tenenbaum (2008). Church: a language for generative models. Uncertainty in Artificial Intelligence 2008.