# Eight-Month-Old Infants' Social Evaluations of Agents Who Act on False Beliefs

**Brandon Woo (bmwoo@g.harvard.edu) and Elizabeth Spelke (spelke@wjh.harvard.edu)**
Department of Psychology, Harvard University, Cambridge, MA 02138
Center for Brains, Minds, and Machines, Cambridge, MA 02139

**Abstract**

Do infants' social evaluations privilege the outcomes of others' actions, or the beliefs underlying those actions? In two experiments, 8-month-old infants viewed a protagonist who sought to grasp one of two toys, each inside a different box, as two other agents observed. Then, while the protagonist was away, the toys exchanged locations, either in the presence or absence of the two other agents. Thus, the agents had either true or false beliefs about the toys' locations. When the protagonist returned, one agent opened the box that now contained the protagonist's desired toy, whereas the other opened the box that previously contained that toy. When agents had true beliefs about the desired toy's location, infants preferred the agent who opened the box containing that toy. When agents had false beliefs about that location, infants instead preferred the agent who opened the opposite box. Thus, infants' social evaluations privilege agents' beliefs.

**Keywords:** theory of mind; social evaluation; cognitive development; infancy

## Introduction

To cooperate, communicate, and interact with others effectively, people cannot just focus on their own mental states or on what they know of reality; people must make sense of others' minds and recognize when others hold false beliefs: representations of the world that differ from their own and from the true state of the world. An understanding of others' minds is critical to moral cognition, enabling adults to distinguish between someone who harms intentionally, and someone who does so accidentally, guided by a false belief (Young et al., 2007). The present experiments aim to examine whether preverbal infants engage in such mentalistic, intention-based evaluation of agents who act on false beliefs.

Decades of research have revealed that young children often struggle in verbal tests of false-belief understanding. Until about 4 years of age, children claim that an agent will look for a desired toy in its current location, even if the agent had no way of knowing it had moved from the location where the agent had last seen it (Wellman et al., 2001; see also, Baron-Cohen et al., 1985; Wimmer & Perner, 1983). If young children fail to understand others' false beliefs, then they may struggle to differentiate intentional and unintentional moral actions, when the latter are guided by false beliefs (Killen et al., 2011).

In contrast to the findings of verbal tests, a body of experiments has provided evidence that young children and toddlers demonstrate sensitivity to others' mental states in nonverbal versions of classic false-belief tests (Buttelmann et al., 2009; Clements & Perner, 1994; Onishi & Baillargeon, 2005; Rhodes & Brandone, 2014; Southgate et al., 2007; see also, Scott & Baillargeon, 2017). In one such experiment, Onishi and Baillargeon found that 15-month-old toddlers looked longer when an agent looked for a desired object in the location where it had moved to in her absence, rather than the location where she had last seen it (i.e., where she false believed the object to be). These findings suggest that toddlers expected the agent to act in a way consistent with her false belief.

There are two reasons, however, to believe that early, nonverbal abilities to reason about false beliefs are fragile at best. First, there have been multiple failures to replicate findings of false-belief understanding in toddlers, both by independent groups (Crivello & Poulin-Dubois, 2018; Powell et al. 2018; Wiesmann et al., 2018; Yott & Poulin-Dubois, 2016; Poulin-Dubois et al., 2018; see also, Baillargeon et al., 2018) and by one of the original groups to report these findings (Kampis et al., 2021). In contrast, independent groups have found consistent evidence for toddlers' ability to reason about states of knowledge and ignorance (see Holland & Phillips, 2020). Thus, evidence for false-belief understanding is more difficult to replicate than is evidence for understanding of other mental states, in nonverbal tests on toddlers. Second, whereas evidence for false-belief understanding has mostly come from toddlers in the second year, there is evidence for an understanding of knowledge and ignorance in infants in the first year (see Phillips et al., 2020, for review). Thus, false-belief understanding may emerge later in development than an understanding of knowledge and ignorance.

Most past work, however, has focused on minimally social contexts in which a single agent acts on inanimate objects for its own benefit. This research therefore has not presented infants with the more strongly social contexts that encourager mental state inferences. A large body of research suggests that infants are sensitive to social contexts at young ages. Infants preferentially look to and reach for agents who help others over agents who hinder others as early as 3 months of age (Hamlin et al., 2007, 2010; Hamlin & Wynn, 2011). Moreover, by late in the first year, infants' social evaluations are sensitive to states of ignorance and knowledge (Hamlin et al., 2013; Woo et al., 2017). Finally, by 15 months of age, toddlers demonstrate sensitivity to the intentions of agents who act on false beliefs (Woo & Spelke, 2022). This research



In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* 

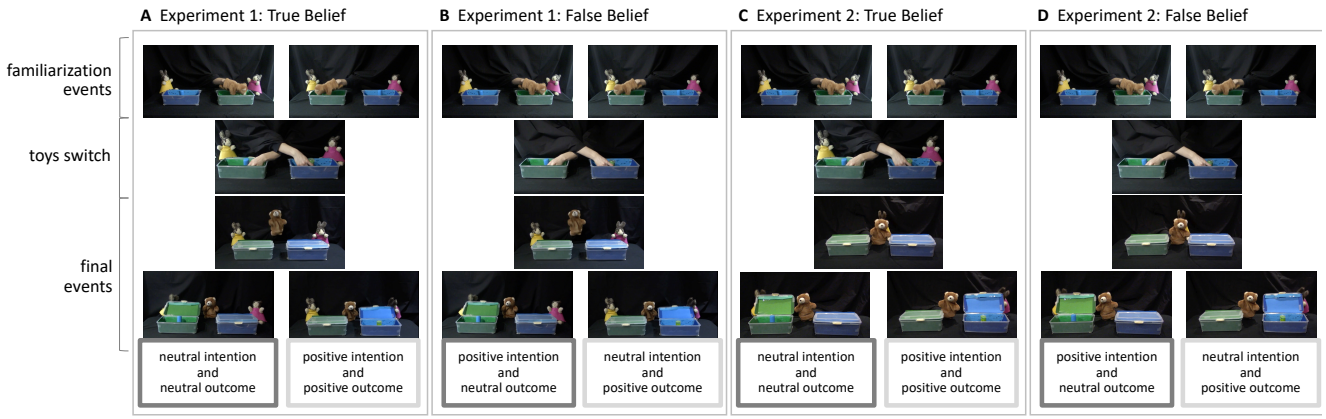


Figure 1: Events presented to infants in Experiments 1 (A, B) and 2 (C, D). In familiarization events, the bear protagonist repeatedly grasped a particular toy in one of two boxes, as two rabbit puppets were present on stage to observe. In the toy-switch event, a pair of hands moved the bear's desired toy to the other box, and put a different toy in the original box that the bear had entered, either as the rabbits were present (A, C) or absent (B, D) to observe. The hands then closed the box. In the final events, each rabbit opened a different box.

stands in contrast to failed attempts to replicate findings of false-belief understanding in nonverbal tests, and suggests that infants and toddlers may be more sensitive to others' false beliefs in contexts in which agents act for others' benefit.

The present experiments test whether 8-month-old infants privilege the intentions of social agents who act on false beliefs. We tested 8-month-old infants, because they demonstrate preferences for helpful individuals in the paradigm that we adapted, when all agents act on true beliefs (Woo & Spelke, 2021). Moreover, infants at this age demonstrate sensitivity to intentions in situations in which agents attempt but fail to help or to harm others (Hamlin, 2013).

We adapted the methods of Woo and Spelke (2022), who studied 15-month-old toddlers, for the present experiments on 8-month-old infants. In two experiments, we familiarized infants to displays depicting a bear protagonist who repeatedly jumped into one of two differently colored, open boxes, each containing a different toy, and they consistently grasped that toy (see Fig. 1). During this action, two rabbits stood on the sides of the stage, witnessing the bear's consistent choice of one toy. Then, while the rabbits were either present or absent, two hands entered the stage, switched the boxes' contents, and closed the boxes. Thus, the bear's desired toy was now in a new box, and a non-desired toy was in the original box that the bear had entered. Importantly, the boxes were opaque, so the rabbits could only have known that the switch had occurred if they were present to observe it. If the rabbits witnessed the switching of the toys (Figs. 1A and 1C), a rational observer could infer that they held true beliefs about the contents of the boxes. If they were absent during the switch (Figs. 1B and 1D), a rational observer would instead infer that they had false beliefs about the contents of the boxes.

In the final events, the bear returned and jumped between the boxes. One rabbit opened the original box, even though the toy inside was now different: a neutral outcome for the bear. The other rabbit opened the box that newly contained the bear's desired toy: a positive outcome for the bear. Infants then saw the two rabbits in a preferential looking choice test.

To assess infants' preferences between the two rabbits, we presented infants with a preference test based on their socially guided looking (after Hamlin & Wynn, 2011; Hamlin et al., 2010). While the two rabbits appeared side by side, a friendly voice called to the baby, encouraging the baby to engage with the rabbits and asking, "Who do you like?" If infants are not sensitive to the false beliefs of social agents, then they may evaluate the agents based on the outcomes of their actions, and look more to the rabbit who causes a positive outcome, regardless of the rabbits' beliefs. In contrast, if infants privilege the intentions of agents acting on false beliefs, then they should look more to the rabbit whose actions were guided by positive intentions. Thus, infants should prefer looking to the rabbit who caused a positive outcome when the rabbits had true beliefs about the outcomes of their actions, and to the rabbit who caused a neutral outcome when the rabbits had false beliefs about the outcomes of their actions.

## Experiment 1

### Method

Hypotheses, methods, and analysis plans for both Experiments 1 and 2 were preregistered on the Open Science Framework. All preregistration documents and stimuli can be found at https://osf.io/2jzpq/.

**Participants** Forty-eight full-term 8-month-old infants contributed data to this experiment (24 girls; mean age: 7.99 months; range = 7;9 to 8;27). An additional 2 participants began the experiment but were excluded due to poor

participant video quality ($n$ = 1) and inattentiveness ($n$ = 1). Experimenters who were unaware of the events infants saw determined exclusions using preregistered criteria. For all studies, participants were tested with informed consent by caregivers.

Our sample size was based on power analyses over pilot data collected with 8-month-old infants in the True Belief Condition, previously collected data with 8-month-old infants in an experiment using methods similar to the True Belief Condition (Woo & Spelke, 2021), and previously collected data with 15-month-old toddlers that inspired the present experiments (Woo & Spelke, 2020).

**Displays** Each infant viewed 6 familiarization events, 1 event in which toys switched positions, 4 final events, followed by a single 30-second social preference test. The first 11 events depicted two opaque boxes (one blue, one green), and two toys (one blue, one green) that were inside the boxes. Events are outlined below (see Fig. 1). In familiarization and the final events, infants saw the video loop four times per event.

Familiarization events began with two rabbits (one wearing a pink shirt, one wearing a yellow shirt) sitting at a stage's rear corners, and two open boxes (one blue, one green), each with a toy of the same color inside. At the start of each event, a bear puppet (the protagonist) jumped onto the stage in between the boxes, and jumped directly into a box to grasp the toy inside.

In all 6 familiarization events, the protagonist always approached and jumped into the same box to grasp the toy inside, demonstrating that it had a preference for that toy. Between familiarization events, the two boxes switched locations. Thus, the box and toy that the protagonist approached appeared alternately on the left and the right.

After familiarization, while the protagonist was away from the stage, infants saw a single toy-switch event in which a pair of hands entered the stage from behind and switched the toys. Thus, the original box that the protagonist had jumped into now contained a different toy, and the other box now contained the toy that the protagonist had consistently grasped. The hands then closed the boxes. In the True Belief Condition (Fig. 1A), the rabbits were present to observe the change of the toys' locations, and could be attributed with knowledge of the switch. In the False Belief Condition (Fig. 1B), the rabbits were absent, and therefore could instead be attributed with ignorance that the switch has happened, and a false belief about the box that contained the desired toy.

In the 4 final events, the two rabbits began sitting at the stage's rear corners, as in familiarization. The boxes were on stage as at the end of the toy-switch event, both closed. At the start of each final event, the protagonist jumped onto the stage at the center, and jumped up and down as though calling for attention. In alternating events, one rabbit moved forward to open the original box that the protagonist had approached, even though the toy inside was now different, and the other rabbit moved forward to open the new box that contained the toy that the protagonist had previously chosen. The social preference test, presented the two rabbits with no boxes during a 30-second period in which a socially engaging voice called to the infant and asked "Who do you like?" A 30-second period has been used in past work probing infants' social evaluations using preferential looking measures (e.g., Hamlin & Wynn, 2011; Hamlin et al., 2010; Woo & Spelke, 2021).

**Procedure** Data collection occurred during the COVID-19 pandemic, and took place over Zoom video calls. Infants sat on their caregivers' laps or in highchairs, and viewed displays on laptops ($n$ = 44), phones ($n$ = 2), a desktop ($n$ = 1), or a tablet ($n$ = 1). Caregivers were instructed to sit quietly and not influence their infants, and to look away from displays in the toy-switch and final events.

We probed infants' evaluations by measuring their preferential looking to the rabbits, following all events. Before presenting the rabbits, we used attention grabbers to obtain reference points for coding. We then recentered each infant's gaze using an attention grabber. Next, the two rabbits appeared on opposite sides of the screen and moved to a prerecorded voice saying "Hi! Look! Who do you like?" three times, once every 10 seconds over a 30-second period. An experimenter, who was unaware of condition and of the events that infants had seen, coded the videos of infants in this looking preference test to determine how much time infants spent looking at each rabbit. Based on these times, we calculated the proportion of time infants spent looking at the rabbit with positive intentions.

A second experimenter, who was unaware of the experimental condition and of the events, coded a randomly selected 25% of infants. For the preference test, the intraclass correlations between the two coders' looking times were 0.98 (95% CI[0.96, 0.99]) for both left- and right-looking.

**Counterbalancing** The following were counterbalanced across infants: the color of the toy and box that the protagonist approached in familiarization, the side of the rabbit with positive intentions throughout events, the order in which the rabbit with positive intentions acted in the final events, and the color of the rabbit with positive intentions.

## Results

**Preregistered Analyses** All reported p-values are two-tailed. In both experiments, for the two conditions, we first calculated the proportion of time infants looked at the rabbit who provided access to the preferred toy during the social preference test. We ran a one-sample $t$-test to determine whether the proportion of time looking at the rabbit who had positive intentions differed from 50% within each condition.

In Experiment 1's True Belief Condition, infants looked more to the rabbit who caused a positive outcome, guided by positive intentions ($\text{mean}_{\text{positive-outcome, positive-intention}}$ % = 55.0%, 95% CI [51.6%, 58.3%], $SD$ = 7.8%, one-sample $t(23) = 3.11$, $p = .004$, $d = 0.63$). In the False Belief Condition, by contrast, infants looked more to the rabbit who opened the original box that the protagonist had jumped into, even though the toy

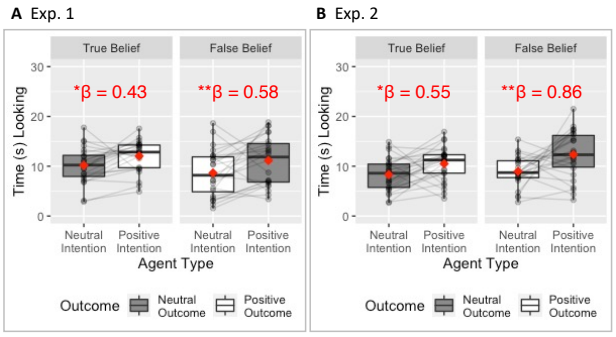


Figure 2: Results in Experiments 1 and 2. Graphs depict the mean time each infant looked to each rabbit by condition in the social looking preference test. Red diamonds indicate means and connected dots indicate data from individual toddlers. Horizontal lines within boxes indicate medians, boxes indicate interquartile ranges, and whiskers indicate 1.5 times the interquartile range. The beta coefficients (β) indicate standardized effect sizes. Across panels, asterisks indicate significant differences (*$p < .05$, **$p < .01$).

inside was now different ($mean_{neutral\text{-}outcome,\ positive\text{-}intention}$ % = 57.4%, 95% CI [51.7%, 63.2%], *SD* = 13.5%, one-sample $t(23) = 2.71$, $p = .012$, $d = 0.55$). Preferences based on outcomes differed significantly between conditions (two-sample $t(36) = 3.90$, $p < .001$, $d = 1.12$).

**Exploratory Analyses** Our preregistered analyses were based on analyses of proportions. In exploratory analyses on raw looking time in the choice test, we found converging results. We ran an exploratory mixed-effects model, in which the dependent variable was time looking at a target; the fixed effects were the outcome caused by a rabbit (Neutral-Outcome = -0.5, Positive-Outcome = 0.5), condition (True Belief = -0.5, False Belief = 0.5), and the interaction; and there was a random intercept for participant ID. Fixed effects were centered.

Whereas neither outcome nor condition alone predicted looking time (*p*s > .228), there was a significant interaction of outcome and condition (β = -1.02, 95% CI of β [-1.59, -0.45], $b = -4.35$, $t(48) = -3.48$, $p = .001$). Posthoc pairwise tests, correcting for multiple comparisons using Holm's method, revealed that infants in the True Belief Condition looked longer to the rabbit causing a positive outcome ($mean_{positive\text{-}outcome}$ = 12.05 s, *SD* = 3.18 s) than to the rabbit causing a neutral outcome ($mean_{neutral\text{-}outcome}$ = 10.17 s, *SD* = 3.59 s) (β = -0.43, $b = -1.87$, $t(50) = -2.07$, $p = .043$), and infants in the False Belief Condition looked longer to the rabbit causing a neutral outcome ($mean_{neutral\text{-}outcome}$ = 11.14 s, *SD* = 4.84 s) than to the rabbit causing a positive outcome ($mean_{positive\text{-}outcome}$ = 8.66 s, *SD* = 4.69 s) (β = 0.58, $b = 2.48$, $t(50) = 2.75$, $p = .008$).

## Discussion

Experiment 1's findings suggest that infants evaluated the rabbits based on the rabbits' intentions, rather than on the outcomes that they caused. Infants looked to the rabbit with positive intentions in both conditions, even when that rabbit had produced a neutral outcome when the rabbits had not observed the switch in the toys, and therefore held a false belief about the contents of the boxes.

To facilitate infants' tracking of the rabbits' actions, the rabbits appeared and acted in constant positions in Experiment 1. On each of the final events, then, the rabbit could be attributed with choosing whether to act, but not as strongly attributed with choosing where to act, given that each rabbit acted on the box that was closest to it. Infants nevertheless formed preferences between the rabbits under these circumstances, providing evidence that the infants were sensitive to the rabbits' choice of whether or not to open a box. In Experiment 2, we replicated the present findings in a situation in which the rabbits more clearly chose which of the two boxes to act upon.

# Experiment 2

## Method

**Participants** We had preregistered a sample size of 48 based on power analyses over the data from Experiment 1. More caregivers responded than we had anticipated, resulting in a sample of 53 full-term 8-month-old infants (24 girls; mean age: 7.88 months; range = 7;9 to 8;27). An additional 6 participants began the experiment but were excluded due to inattentiveness ($n = 4$), equipment failure ($n = 1$), and caregiver influence ($n = 1$).

**Displays, Procedures, and Counterbalancing** Displays, procedures, and counterbalancing were the same in Experiment 2 as those of Experiment 1, with one exception: In the final events, only the protagonist and one of the rabbits were present (see Figs. 1C and 1D). On alternating events, each of the two rabbits now began positioned behind the protagonist, and thus, more clearly chose where to act and which box to act upon. As in Experiment 1, one rabbit consistently opened the box containing the desired toy, and the other rabbit consistently opened the other box.

In Experiment 2, infants viewed displays on laptops ($n = 46$), desktops ($n = 3$), tablets ($n = 2$), or phones ($n = 2$).

In addition to a primary experimenter, a second experimenter coded a randomly chosen 25% of infants' preference tests. The intraclass correlations between the two coders' looking times were 0.90 (95% CI[0.73, 0.97]) and 0.91 (95% CI [0.75, 0.97] for left- and right-looking, respectively.

## Results

**Preregistered Analyses** In Experiment 2's True Belief Condition, infants looked more to the rabbit who caused a positive outcome ($mean_{positive\text{-}outcome,\ positive\text{-}intention}$ % = 56.2%, 95% CI [51.2%, 61.1%], *SD* = 12.4%, one-sample $t(26) = 2.59$, $p = .015$, $d = 0.63$). In the False Belief Condition, by contrast, infants looked more to the rabbit who opened the original box that the protagonist had jumped into, even

though the toy inside was now different ($mean_{neutral\text{-}outcome,\ positive\text{-}intention}$ % = 56.7%, 95% CI [51.5%, 61.9%], *SD* = 12.8%, one-sample $t(25) = 2.67$, $p = .013$, $d = 0.52$). Preferences based on outcomes differed significantly between conditions (two-sample $t(50) = 3.72$, $p < .001$, $d = 1.02$).

**Exploratory Analyses** As in Experiment 1, we examined whether infants looked longer at the rabbit with positive intentions, as in Experiment 1. The model specifications were the same. Whereas neither outcome nor condition alone predicted looking time (*p*s > .084), there was a significant interaction of outcome and condition (β = -1.41, 95% CI of β [-2.11, -0.73], $b = -5.53$, $t(53) = -4.05$, $p < .001$). Posthoc pairwise tests, correcting for multiple comparisons using Holm's method, revealed that infants in the True Belief Condition looked longer to the rabbit causing a positive outcome ($mean_{positive\text{-}outcome}$ = 10.50 s, *SD* = 3.35 s) than to the rabbit causing a neutral outcome ($mean_{neutral\text{-}outcome}$ = 8.31 s, *SD* = 3.27 s) (β = -0.55, $b = -1.87$, $t(55) = -2.18$, $p = .029$), but that infants in the False Belief Condition looked longer to the rabbit causing a neutral outcome ($mean_{neutral\text{-}outcome}$ = 16.18 s, *SD* = 4.69 s) than to the rabbit causing a positive outcome ($mean_{positive\text{-}outcome}$ = 11.09 s, *SD* = 2.99s) (β = 0.86, $b = 2.48$, $t(55) = 3.36$, $p = .018$).

### Discussion

Experiment 2's findings replicated those of Experiment 1, in a situation in which rabbits more clearly chose which box to act on. These findings again are consistent with intention-based evaluations based on false-belief inferences.

## General Discussion

In two experiments, 8-month-old infants inferred the beliefs of two agents about the location of a desired object, based on whether agents were present or absent to observe a change in the state of the world. Infants engaged in intention-based evaluations of the agents, based on their inferred beliefs. Specifically, infants preferred (i) an agent who produced a neutral outcome when agents acted on false beliefs about the outcomes of their actions, and (ii) an agent who produced a positive outcome when agents acted on true beliefs about the outcomes of their actions. In both conditions, the preferred agent demonstrated an intention to produce the bear's desired outcome. Thus, infants privileged intentions over outcomes in their evaluations, and demonstrated sensitivity to the intentions and beliefs underlying the actions of the two social agents.

The present findings are consistent with research on 15-month-old toddlers (Woo & Spelke, 2022), on which the present paradigms were based, but they are opposed by a large body of research finding that young children focus on outcomes, rather than intentions, in verbal tasks probing their moral judgments. In contrast to the latter research, the present findings contribute to a growing body of evidence that infants and young children are sensitive to others' intentions in social contexts, in which an agent's actions have potential consequences for other agents (Hamlin, 2013; Hamlin et al., 2013; Kanakogi et al., 2017; Woo et al., 2017). Here, the rabbit with positive intentions could be seen as wishing to help the bear, to match the bear's preference, or both.

The present findings and those of Woo and Spelke (2020) also stand in contrast to recent failures to replicate evidence of toddlers' sensitivity to agents' beliefs in contexts in which agents act for their own benefit (see Poulin-Dubois et al., 2018). Moreover, whereas most positive evidence for sensitivity to agents' beliefs has come from studies of toddlers in the second year, the present evidence is based on studies of infants in the first year. Researchers have proposed that sensitivity to agents' beliefs may emerge in the second year, long after the emergence of sensitivity to agents' states of knowledge and ignorance (Phillips et al., 2021). The present findings challenge this idea, and suggest that sensitivity to beliefs emerges earlier in infancy, when agents' beliefs have social consequences.

Why might infants and toddlers be more sensitive to other agents' beliefs when studies probe social evaluations, rather than predictions? One possibility is that infants may focus on the beliefs and intentions of agents whose actions have social consequences, because their beliefs can shed light on the likelihood that an agent will later cooperate or act generously. We look forward to research that more directly tests this possibility by comparing infants' and toddlers' mental state reasoning in contexts in which agents act either for their own or for others' benefit.

A second possibility is that demands may differ between studies probing expectations vs. social evaluations. Studies probing expectations depend on participants forming a prediction about an agent's future behavior. By contrast, studies probing social evaluations do not involve such a challenge. We look forward to research that more directly probes how task demands relate to early, implicit false-belief understanding.

In sum, in two experiments, infants considered the beliefs of two agents. Infants viewed these agents as having representations of the world that modulated their intentions, and infants formed preferences for agents who had positive intentions, regardless of the outcomes that those agents caused. Such an early-emerging sensitivity to other agents' beliefs and intentions may support children's navigation of the social world.

## Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines, funded by National Science Foundation STC award CCF-1231216. We thank the families who volunteered to participate, and we thank Cameron Calderwood for research assistance.

## References

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112-124.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition*, *21*(1), 37-46.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337-342.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377-395.

Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development*, *46*, 51-57.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451-474.

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

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30-39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559.

Hamlin, J. K., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923-929.

Holland, C., & Phillips, J. (2020). A theoretically driven meta-analysis of implicit theory of mind studies: The role of factivity. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2020). A two-lab direct replication attempt of Southgate, Senju, & Csibra (2007). *Royal Society Open Science*, *8*(8), 210190.

Kanakogi, Y., Inoue, Y., Matsuda, G., Butler, D., Hiraki, K., & Myowa-Yamakoshi, M. (2017). Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour*, *1*(2), 1-7.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197-215.

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

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

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., ... & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet–A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, *48*, 302-315.

Rhodes, M., & Brandone, A. C. (2014). Three-year-olds' theories of mind in actions and words. *Frontiers in Psychology*, *5*, 263.

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

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

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

Wiesmann, C. G., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false belief-related action anticipation. *Cognitive Development*, *46*, 58-68.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103-128.

Woo, B. M., & Spelke, E. S. (2021). *Infants' and toddlers' evaluations of helpers depend on their understanding of action goals*. PsyArxiv. https://doi.org/10.31234/osf.io/mtprn

Woo, B. M., & Spelke, E. S. (2022). *Toddlers' social evaluations of agents who act on false beliefs*. PsyArxiv. https://doi.org/10.31234/osf.io/eczgp

Woo, B. M., Steckler, C. M., Le, D. T., & Hamlin, J. K. (2017). Social evaluation of intentional, truly accidental, and negligently accidental helpers and harmers by 10-month-old infants. *Cognition*, *168*, 154-163.

Yott, J., & Poulin-Dubois, D. (2016). Are infants' theory-of-mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development*, *17*(5), 683-698.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235-8240.