Toddlers' Social Evaluations of Agents Who Act on False Beliefs

Brandon M. Woo and Elizabeth S. Spelke

Department of Psychology, Harvard University, Cambridge, MA 02138

The Center for Brains, Minds, and Machines, Cambridge MA, 02139

This manuscript is in press at *Developmental Science*.

Correspondence should be addressed to: Brandon Woo (bmwoo@g.harvard.edu)

ORCID ID

Brandon M. Woo: https://orcid.org/0000-0002-8639-2919

Elizabeth S. Spelke: https://orcid.org/0000-0002-6925-3618

Data Availability Statement

All experiments were formally preregistered. All deidentified data are hosted on the Open Science Framework at https://osf.io/9pd6q/?view_only=3a820835ca804463ac771db8a5f112ed.

Funding Statement

This work was supported by the Center for Brains, Minds, and Machines, funded by National Science Foundation STC Award CCF-1231216; and by a Social Sciences and Humanities Research Council Doctoral Fellowship under award 752-2020-0474.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgements

We thank the families who participated in these experiments, the Cambridge Writing Group and Tomer Ullman for feedback, Bill Pepe, Larisa Shrestha, Will Adams,

Kexin Que, and Yuman Li for research assistance, Hyowon Gweon and the Stanford Social Learning Lab for sharing protocols to facilitate online testing, and Benedek Kurdi for statistics consulting.

Abstract

Mature social evaluations privilege agents' intentions over the outcomes of their actions, but young children often privilege outcomes over intentions in verbal tasks probing their social evaluations. In three experiments ($N = 118$), we probed the development of intention-based social evaluation and mental state reasoning using nonverbal methods with 15-month-old toddlers. Toddlers viewed scenarios depicting a protagonist who sought to obtain one of two toys, each inside a different box, as two other agents observed. Then, the boxes' contents were switched in the absence of the protagonist and either in the presence or the absence of the other agents. When the protagonist returned, one agent opened the box containing the protagonist's desired toy (a positive outcome), and the other opened the other box (a neutral outcome). When both agents had observed the toys move to their current locations, the toddlers preferred the agent who opened the box containing the desired toy. In contrast, when the agents had not seen the toys move and therefore should have expected the desired toy's location to be unchanged, the toddlers preferred the agent who opened the box that no longer contained the desired toy. Thus, the toddlers preferred the agent who intended to make the protagonist's desired toy accessible, even when its action, guided by a false belief concerning that toy's location, did not produce a positive outcome. Well before children connect beliefs to social behavior in verbal tasks, toddlers engage in intention-based evaluations of social agents with false beliefs.

Keywords: theory of mind, false-belief understanding, social evaluation, intention, cognitive development, online testing

Toddlers' Social Evaluations of Agents Who Act on False Beliefs

A man sells his watch to buy combs for his wife, not knowing that she has sold her hair (Henry, 2012). A child recycles a classmate's bag, not knowing that the bag contained a desired cupcake (Killen et al., 2011). In these cases, the man's and the child's false beliefs render the man's gift useless and the child's action hurtful, but both individuals can be inferred to have helpful intentions. When adults evaluate agents within such scenarios, they privilege the agents' intentions over the outcomes of their social actions, even when intentions depend on false beliefs (Cushman et al., 2006; Knobe, 2005; Young et al., 2007). The development of false-belief understanding is a topic of intense research and debate (Phillips et al., 2020; Poulin-Dubois et al., 2018; Scott & Baillargeon, 2017). Here we ask whether toddlers, like adults, favor agents who intend to produce positive outcomes for other agents, but fail to achieve those outcomes because their actions are guided by false beliefs.

Children's Developing Theory of Mind

Decades of research have found that children under 4 years of age verbally judge agents who cause negative outcomes harshly, regardless of whether the agents intended to cause harm or falsely believed that their actions would be helpful (Cushman et al., 2013; Margoni & Surian, 2016; Piaget, 1965; Yuill & Perner, 1988; Zelazo et al., 1996). These findings accord with evidence that young children struggle to reason about beliefs that oppose what they know of reality (Baron-Cohen et al., 1985; Wellman et al., 2001; Wimmer & Perner, 1983). Social evaluations appear to hinge first on concrete outcomes and later on abstract intentions, because young children fail to consider the mistaken beliefs that lead others' intentions to conflict with the outcomes of their actions

(Astington, 2004; Killen et al., 2011; Smetana et al., 2014). Although an ability to understand others' intentions emerges in infancy (Sodian et al., 2020; Woodward, 2009), when others' intentions conflict with the outcomes of their actions, young children appear to focus on the outcomes of others' actions.

Whereas young children struggle to evaluate unintended actions that are guided by false beliefs, toddlers have succeeded at nonverbal versions of one classic false-belief scenario, in which an agent acts to recover a hidden object that had changed its location while out of the agent's view (D. Buttelmann et al., 2009; Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017; Southgate et al., 2007). Diverse failures to replicate these findings, however, suggest that toddlers' understanding of false beliefs is fragile at best (Holland & Phillips, 2020; Phillips et al., 2020; Poulin-Dubois et al., 2018; Powell et al., 2018). Yet, most studies probing toddlers' understanding of false beliefs have focused on agents who act for their own benefit. For example, in the displays of Onishi and Baillargeon (2005), a stranger seeks an inanimate object. In such minimally social contexts, even adults rarely have reason to care about the beliefs that guide unknown agents' actions. In contrast, beliefs are central to understanding social actions. The same act of throwing away a bag containing someone's lunch could be cruel if one had done so knowingly, or an honest mistake if one had believed the bag to be empty. Studies suggest that young children demonstrate enhanced capacities for mental state reasoning in more strongly social contexts (Asaba & Gweon, 2019; Tsoi et al., 2020, 2021), raising the possibility that toddlers can track agents' beliefs when the agents' actions have consequences for their social partners.

Early Capacities for Evaluating Social Agents

Beginning with Hamlin et al. (2007), studies have found that infants and toddlers evaluate agents based on their social actions (Margoni & Surian, 2018; Powell & Spelke, 2018; Thomas & Sarnecka, 2019). For example, 3- and 6-month-old infants look at and reach for, respectively, an agent who had helped a protagonist climb a hill over one who had pushed the protagonist down the hill (Hamlin et al., 2007, 2010). Despite some failures to replicate these findings (Salvadori et al., 2015; Schlingloff et al., 2020), a meta-analysis of experiments probing early social evaluations found that infants and toddlers demonstrate a significant preference for prosocial over antisocial agents (Margoni & Surian, 2018). Since this meta-analysis, 13 additional studies have provided evidence that infants and toddlers evaluate agents based on their social actions (see Woo et al., 2022), consistent with the hypothesis that infants and toddlers prefer agents who help others to achieve their goals. Margoni and Shepperd (2020) have found, moreover, that the failed replications in this literature are consistent with the expected levels of error (i.e., false negatives) that occur when many samples are drawn from a population whose evaluations show a true preference for prosocial agents. The evidence that infants and toddlers engage in social evaluation, then, appears more robust than the evidence that toddlers implicitly represent others' false beliefs. Still, there has been great debate concerning the basis of early social evaluation (Hamlin, 2013a; Powell, 2021). Do infants and toddlers prefer helpful agents because their actions result in positive outcomes or because they possess positive intentions?

A growing body of research provides evidence that infants and toddlers evaluate others in accord with the intentions underlying their social actions, in situations where

agents hold no false beliefs (Geraci et al., 2022; Hamlin, 2013b; Kanakogi et al., 2017). In one experiment (Hamlin, Ullman, et al., 2013), 10-month-old infants selectively reached for an agent who provided a protagonist with its desired toy only when the agent had previously seen the protagonist selectively approach that toy, demonstrating its preference. If instead the protagonist expressed its preference in the agent's absence, infants did not favor the agent who produced the desired toy. These findings provide evidence that by 10 months of age, infants' evaluations of helping are not solely outcome-based and depend on the agent's state of knowledge or ignorance. Further research on 10-month-old infants provides additional evidence for early intention-based evaluation of knowledgeable and ignorant agents (Woo et al., 2017).

Although states of knowledge and ignorance are related to states of true and false beliefs, inferences of the latter present further challenges. When one attributes a false belief to another agent, one assigns a representation of the world that differs from both the true state of the world and the representations held by the self (Hogrefe et al., 1986; Tomasello, 2018). Research with children (Hogrefe et al., 1986; Wellman & Liu, 2004), nonhuman primates (Kaminski et al., 2008; Marticorena et al., 2011), and adults (Apperly et al., 2008) (see Phillips et al., 2020) provides evidence that reasoning about false beliefs poses greater difficulties than does reasoning about knowledge and ignorance. Although infants and toddlers have demonstrated abilities to evaluate others based on their intentions, it remains an open question whether infants and toddlers evaluate social agents based on their beliefs, when those beliefs are at odds with reality and with the children's own beliefs. Here we test for this ability at 15 months of age: the age of

participants in past research probing early false-belief understanding (Onishi & Baillargeon, 2005).

Research Overview

Three experiments tested 15-month-old toddlers' evaluations of two social agents who both either saw or did not see a protagonist's desired toy move to a new location. These experiments adapt methods by Woo and Spelke (2022), who probed toddlers' understanding of the goals of a bear protagonist who, with help from two rabbit agents, opened a closed box and then grasped a toy inside the box. By late in the first year, infants infer that the goal of such an agent is to obtain the toy, rather than to open the box (Woodward & Sommerville, 2000). After the bear had obtained the toy several times, two hands emerged and moved the desired toy to a different box; in this paradigm, the rabbits were present to observe this movement. Following this movement, the bear appeared, and each rabbit opened a different box: either the current or the former box that had contained the desired toy. When later presented with the two rabbits, 15-month-old toddlers both preferentially reached for (in-person) and looked at (over video calls) the rabbit who had opened the box containing the desired toy. Given these positive findings, we adapted this paradigm to probe toddlers' evaluations of agents who act on false beliefs.

We introduced one key change: We manipulated whether the rabbits observed the bear's desired toy move to a different box. This manipulation mirrors those used in studies probing false-belief reasoning (Krupenye et al., 2016; Onishi & Baillargeon, 2005; Southgate et al., 2007; Wellman et al., 2001). When neither rabbit agent had witnessed the desired toy move, would toddlers prefer the agent who directed the

protagonist to the location where it had last seen that toy (i.e., where both agents falsely believed it to be), or the agent who directed the protagonist to the toy's current location?

In Experiments 1-3, toddlers were familiarized to videotaped puppet shows involving a bear protagonist and two rabbit agents. In familiarization, the bear approached one of two different boxes, each containing a different toy, and grasped the toy inside that box in the rabbits' presence (Fig. 1A). Then the bear left the scene and two hands emerged, switched the boxes' contents, and closed the boxes, leaving the bear's desired toy in the other box. In Experiment 3, the two hands returned to the scene and restored the toys to their original locations.

In each experiment, we manipulated the rabbits' perceptual access to the changes in the locations of the toys. In the True Belief Conditions (Fig. 2A and 2C), the rabbits were present and observed the change in the toys' locations; a rational observer should infer that both rabbits knew which box contained the desired toy (Fig. 1B). In the False Belief Conditions (Fig. 2B and 2D), the rabbits were absent during this change (or during the final change in Experiment 3); as in classic false-belief tests reasoning, a rational observer would not expect the rabbits to update their representations of the toys' locations (Fig. 1B). Following the changes, the bear returned, and in the final events, each rabbit opened a different box. In each experiment, one rabbit gave the bear access to its desired toy and the other did not. In the False Belief Conditions, however, the rabbits' actions were unintentional, as they did not see the desired toy move to its current location. In a fourth experiment, we presented 6- and 7-year-old children with these events and probed their verbal understanding of the rabbits' beliefs and intentions.

To determine whether toddlers evaluated the rabbits based on their beliefs and the intentions behind their actions, we measured toddlers' selective reaching or looking towards the agents: the primary measures used in research on early social evaluations. If 15-month-old toddlers both evaluate social agents based on their intentions and infer others' beliefs based on what they have and have not seen, then toddlers should reach for and look to the rabbit with helpful intentions: the rabbit who opened the box containing the bear's desired toy when the rabbits had seen the toy move there (the True Belief Conditions), and the rabbit who opened the box where it had last seen the bear's desired toy when the rabbits had not seen the toy move to its current location (the False Belief Conditions). If toddlers instead evaluate social agents based on the outcomes that they cause, or are not sensitive to others' beliefs, then toddlers should reach for and look to the rabbit that opened the box containing the desired toy.

Experiment 1

Experiment 1 investigated whether 15-month-old toddlers preferentially reached for an agent with helpful intentions over one with unhelpful intentions, when two agents acted with either true or false beliefs concerning the location of the protagonist's desired toy.

We presented toddlers with videotaped events depicting a puppet stage with two distinctively colored boxes and three puppets: one bear protagonist and two rabbit agents. For the True Belief Condition, the boxes were transparent and each contained a toy of the same color as its box; for the False Belief Condition, the boxes were opaque—their contents could not be seen. The two boxes appeared in alternation on each of the stage's two sides during familiarization; the bear appeared in the center, and each rabbit

consistently appeared on one of the two sides. In both conditions, the bear repeatedly approached and unsuccessfully attempted to open the same colored box by itself; the bear succeeded in opening the box with help of the rabbit closer to that box. Once the box was open, the bear grasped the toy inside. From such preferential approach behavior, regardless of where an object is, infants attribute object-directed goals to others (Luo, 2011; Woo et al., 2021). Thus, the toddlers should have inferred that the bear desired the colored toy that it grasped.

After this familiarization, the scene reappeared without the bear, and two hands entered the stage, opened the two boxes, and exchanged the toys that they contained. In the True Belief Condition, the rabbits were present to observe the desired toy move to the opposite box. In the False Belief Condition, the rabbits were absent when the desired toy moved. For both conditions' final events, the bear returned, and the two rabbits acted in alternation, with each opening a different box. To obtain an exploratory measure of toddlers' interest during events, we compared toddlers' looking times to the two final events. To measure toddlers' evaluations of the two rabbits, and their weighting of the rabbits' intentions vs. the outcomes that the rabbits caused, a social reaching preference test followed.

Method

Methods and analyses for all experiments were preregistered on the Open Science Framework (OSF) at https://osf.io/9pd6q/?view_only=3a820835ca804463ac771db8a5f112ed. Details of power analyses and sample size justifications are included in the preregistrations and SI. Stimuli, data, and code are available on the OSF.

Participants. Forty-six full-term 15-month-old toddlers contributed data (22 randomly assigned to the True Belief Condition, 24 to the False Belief Condition; mean age = 15.06 months; range = 14;10-15;18; 23 girls, 23 boys; see SI for additional demographic details for Experiments 1-4). Fifteen additional participants were excluded due to fussiness (n = 6), failing to reach for a puppet (n = 3), inattentiveness (n = 2), parental interference (n = 2), or procedural error (n = 2). For all experiments, experimenters who were unaware of the events that children saw, the experimental condition, and the role of each puppet determined exclusions using preregistered criteria.

For all experiments, participants were tested with their caregivers' informed consent, and study protocols were approved by [university IRB; hidden for review].

Displays. Each toddler viewed 6 familiarization events, 1 event in which toys switched locations, and 4 final events, for a total of 11 events. Studies probing early social evaluations have presented toddlers around this age with a similar number of events (Hamlin, Mahajan, et al., 2013; Woo & Spelke, 2020).

All events took place on a puppet stage containing 2 boxes (one blue, one green) and 2 toys (one blue, one green, matching the boxes) inside the boxes. The boxes were transparent in the True Belief Condition and opaque in the False Belief Condition.

All familiarization events began with two rabbit puppets (one wearing a pink shirt, one wearing yellow) sitting at the puppet stage's rear corners. Each rabbit remained on the same side of the stage throughout the events, so that the toddlers could better track them by using either their colors or positions. In the 6 familiarization events, the bear protagonist appeared and repeatedly tried and failed to open one box by itself; this box appeared on each of the two sides of the stage in alternating events. Following the bear's

failures to open the box, each rabbit joined the bear on alternating trials, when that box was located on that rabbit's side; together they opened that box, allowing the bear to grasp the toy inside.

Throughout familiarization, the bear only approached and tried to open the box containing its desired toy, and the rabbits only opened that box when the bear attempted to open it; the rabbits refrained from opening the other box, which contained the undesired toy. Between familiarization events, the boxes switched locations. Thus, the box that the bear approached appeared alternately on the left and right.

After familiarization, while the bear was absent from stage, two hands switched the boxes' contents, such that the original box that the bear had tried to open now contained a new toy, and the other box contained the toy that the bear had grasped in familiarization.

We manipulated the rabbits' perceptual access when the toys exchanged locations. In the True Belief Condition (Fig. 2A), the two rabbits were present and observed the change. In the False Belief Condition (Fig. 2B), the two rabbits were absent and did not observe the change.

Following the change in the toys' locations, the bear (and, in the False Belief Condition, the rabbits) returned in the final event phase, and the bear jumped between the two closed boxes as if calling for attention. In 4 alternating final events, one rabbit opened the original box that previously contained the desired toy (but now contained the undesired toy), and the other rabbit opened the other box, now containing the desired toy. Thus, one rabbit rendered the undesired toy available (a neutral outcome for the bear), and the other rendered the desired toy available (a positive outcome). Throughout the

False Belief Condition's final events, the boxes' lids blocked the rabbits' view of the objects (Fig. S1), such that the rabbits remained ignorant of the change in the toys' locations. All actions then ceased while toddlers' looking time was recorded.

Procedure. Each toddler sat on their caregiver's lap in the lab before a 102-cm by 132-cm LCD projector screen. Caregivers were instructed to close their eyes and not influence their toddlers. We required that the toddlers look at the screen while the bear struggled to open a box, a rabbit opened the box, the bear grasped the toy, and the toys switched locations. If a toddler did not see one or more of these critical parts, we repeated the full event (see SI for details, for Experiments 1-3).

After a toddler had watched all events, an experimenter, who was unaware of condition and the events that the toddler had seen, presented the toddler with the two rabbit puppets in a social reaching preference test. The caregiver turned 90° to the left so that they no longer faced the screen. The experimenter then kneeled in front of the toddler and held the puppets approximately 30 cm apart, initially out of each toddler's reach. The toddlers were required to look at both rabbits before looking back to the experimenter; the rabbits were then moved within reach and the experimenter said, "Who do you like?" The experimenter judged which puppet a toddler contacted by means of a visually guided reach. A second researcher, who was unaware of condition and the events, judged which rabbit a toddler looked to and touched first in this test. There was 100% agreement between the two researchers. For all experiments, see SI for counterbalancing information and analyses of preferences in relation to counterbalanced variables.

Coding of Interest in Familiarization and Final Events. To measure toddlers' interest in the events, we used a toddler-controlled looking time procedure. Each

familiarization and final trial ended with a pause, during which looking time data were coded online using Xhab64 (Pinto, 1996) software until a toddler looked away for 2 consecutive seconds or until 30 seconds elapsed. The coder watched the toddlers through a live video in a separate room, could not hear or see events, and was unaware of condition and the events each toddler had seen. A second coder, who was unaware of condition and the events, coded the final events of a randomly selected 25% of toddlers using jHab (Casstevens, 2007) software. The correlation between the two coders' looking times was 0.99.

Results

The toddlers in the True Belief Condition reached to the rabbit that opened the box containing the desired toy, producing the positive outcome for the bear, consistent with helpful intentions (17/22 toddlers, binomial $p = .016$, relative risk = 1.54). In contrast, the toddlers in the False Belief Condition reached to the rabbit that opened the other box, guided by helpful intentions but producing no positive outcome (19/24 toddlers, binomial $p = .006$, relative risk = 1.58) (Fig. 3A). Preferences differed significantly between conditions based on outcome ($\chi^2(1) = 12.47$, $p < .001$, Wald's odds ratio = 12.92). In both conditions, the toddlers looked equally at the final events in which the two rabbits acted on different boxes (see SI for exploratory analyses): Toddlers appeared equally interested in the actions of the two rabbits during the final events.

Discussion

Experiment 1's findings suggest that toddlers evaluated the rabbits. Their selective reaching for the rabbit with helpful intentions cannot be attributed to greater interest in that rabbit's actions in final events, because they showed no such increased

looking to the final events involving that rabbit. Moreover, the findings of the False Belief Condition suggested that toddlers' social evaluations were based on social intentions, rather than on the outcomes caused.

Because the boxes were opaque only in the False Belief Condition, however, an alternative explanation for that condition's findings should be considered: The toddlers may have struggled to understand that the bear attempted to open the box to grasp the toy inside. If so, the toddlers might have focused on the opaque boxes rather than on their contents, expecting the bear to want to open the same box as before. To address this explanation, Experiment 2 presented toddlers with opaque boxes in both the True and False Belief Conditions. If toddlers fail to view the toy as the bear's goal when boxes are opaque, then they should favor the rabbit that opened the original box, giving access to the undesired toy, in both conditions. In contrast, if toddlers correctly inferred the bear's goal and the rabbits' intentions, then Experiment 2's findings should accord with those of Experiment 1.

Experiment 2

Experiment 2's method was the same as that of Experiment 1, except as follows. First, the experiment was conducted during the COVID-19 pandemic when in-person testing ceased; participants were therefore tested in their homes, using Zoom video conferencing. Second, the boxes were opaque in both conditions: The conditions differed only in the presence or absence of the rabbits when the two toys were moved to the opposite boxes. We therefore included a pre-familiarization event at the start of the session for both conditions, presenting the two open boxes with the toys inside. This pre-familiarization established that each box contained a different toy, seen by the bear, the

rabbits, and the toddlers. Third, to reduce fussiness related to the length of events, all trials ended after a fixed duration rather than continuing until toddlers looked away.

The main difference in method was a change in our outcome measure. Because a reaching preference test could not be administered remotely, we assessed toddlers' evaluations by means of a social visual preference test: As the rabbits appeared side by side, we measured toddlers' selective looking at each rabbit, while presenting the same prompt as in Experiment 1's reaching test ("Who do you like?"). In lab-based experiments, similar methods have been used to assess infants' and toddlers' evaluations of agents (Colomer et al., 2020; Geraci et al., 2022; Hamlin et al., 2010; Kinzler et al., 2007; Powell & Spelke, 2018), with findings that are consistent with those of studies using reaching measures. Research conducted via video conferencing has used this visual preference method to probe early social evaluations (in situations that do not involve false beliefs), and found that looking and reaching measures converge in infants and toddlers (Woo & Spelke, 2022) (as reviewed above). Other research has used visual preference methods to probe toddlers' understanding of emotion, replicating in-lab findings (Smith-Flores et al., 2022). Across these in-lab and video-conferencing-based experiments, participants have not looked longer to events in which agents engaged in prosocial actions. Thus, infants' and toddlers' visual preferences appear not to be based on greater interest in events in which prosocial agents acted. Nevertheless, as in Experiment 1, we measured toddlers' looking to the final events, to measure their interest in each action.

Methods

Participants. Forty-eight full-term 15-month-old toddlers contributed data (24 in each condition; mean age = 14.91 months; range = 14;10-15;20; 26 girls, 22 boys). Four additional participants were excluded due to inattentiveness (n = 2), fussiness (n = 1), or equipment failure (n = 1).

Displays. Displays were identical to Experiment 1's displays, except as follows. First, the boxes were opaque in both conditions (Figs. 2B and 2C). Second, the experiment began with a pre-familiarization event in which the two boxes were open with the toys inside. Third, we made minor changes to the events to better engage toddlers' attention (see SI for full details). Fourth, to compensate for the potentially decreased attentiveness of toddlers to events on computer screens, rather than in the lab, the final events were not presented using a toddler-controlled looking time procedure. Instead, each final event paused for 2 seconds after a rabbit opened a box, and the video was looped three additional times, with each toddler only being required to watch the events in one of the four loops of an event.

Procedures: Social Visual Preference Test. Because we could not reliably elicit or assess reaching towards puppets by video conference, we probed toddlers' evaluations by measuring their preferential looking to the rabbits. The two rabbits appeared on opposite sides of the screen and moved to an experimenter's prerecorded voice saying "Hi! Look! Who do you like?" thrice, once every 10 seconds over a 30-second period. An experimenter, who was unaware of condition and the events that toddlers had seen, coded the videos to determine how much time a toddler looked at each of the rabbits. We

calculated the proportion of time that a toddler looked at the rabbit with positive intentions.

A second experimenter, who was unaware of experimental condition and of the events, coded the preference tests. For the preference test, the correlations between the two coders' looking times were 0.93 and 0.97 for left- and right-looking, respectively.

Procedures: Coding of Final Events. To measure each toddler's interest in the final events, looking time was coded offline using jHab (Casstevens, 2007). For trials in which a toddler saw a rabbit open a box in the first loop of the video, a coder coded looking behavior from the moment that the box opened until the toddler had looked away for 2 consecutive seconds or until 4 loops of the video had played. The coder was unaware of the events that each toddler had seen and of the experimental condition. A second experimenter coded the final events of a randomly selected 25% of toddlers. For the final events, the correlation between the two coders' looking times was 0.97.

Results

Whereas the toddlers in the True Belief Condition looked more to the rabbit that produced the intended positive outcome ($\text{mean}_{\text{positive-outcome, helpful-intention \%}}$ = 58.2%, 95% CI [51.9%, 64.5%], *SD* = 14.9%, one-sample $t(23) = 2.71$, $p = .012$, $d = 0.55$), the toddlers in the False Belief Condition looked more to the rabbit whose helpful intentions *failed* to produce the positive outcome ($\text{mean}_{\text{neutral-outcome, helpful-intention \%}}$ = 57.0%, 95% CI [50.8%, 63.2%], *SD* = 14.6%, one-sample $t(23) = 2.36$, $p = .026$, $d = 0.48$) (Fig. 3B). Preferences based on outcomes again differed significantly between conditions (two-sample $t(45) = 3.59$, $p < .001$, $d = 1.03$). (Exploratory analyses on raw looking time in the preference test converged with these analyses; see SI.) In contrast, the toddlers in both

conditions looked equally at the final events, in which rabbits acted on different boxes (see SI).

Discussion

In Experiment 2, the toddlers looked preferentially at the social agent with helpful intentions, regardless of the outcomes of its actions. These findings replicate those of Experiment 1, with a stricter design and a different measure. Again, they cannot be explained by different interest in final events in which a rabbit with helpful intentions acted.

Although Experiments 1 and 2's findings are consistent with intention-based evaluations based on false-belief inferences, they have two limitations. First, past research has found that infants prefer agents who imitate other characters by directing their actions to the objects that another character has acted on (Powell & Spelke, 2018). A preference for imitators cannot account for toddlers' performance in the present True Belief Conditions, because the rabbit with helpful intentions acted on a different box from the box that the bear had acted upon. It is possible, however, that the toddlers struggled to reason about the rabbits' intentions in the False Belief Conditions, because reasoning about false beliefs is more difficult than reasoning about true beliefs. If toddlers failed to track the two rabbits' false beliefs, they might have based their social preferences on the similarity of each rabbit's action to the bear's action in familiarization. Such a choice would lead them to favor the rabbit with helpful intentions in the False Belief Condition, not because its intentions were helpful, but because it imitated the bear's action by opening the original box that the bear had attempted to open.

Second, past research has investigated infants' and children's understanding of an agent's beliefs primarily by focusing on a single agent that is faced with a choice between acting on different objects at different locations: For example, should Sally look for her ball in the box on the right or in the basket on the left (Baron-Cohen et al., 1985)? Experiments 1 and 2, in contrast, focused on two agents' that are faced with a choice between acting, or not acting, at a single location containing a single object. Although toddlers formed consistent preferences for the rabbit with helpful intentions, reflected in its choice of *when* to act, these experiments leave open the question whether young children can infer agents' beliefs in the better-studied situation in which an agent chooses *where* to act and *what object* to act upon.

Experiment 3

To address the limitations of Experiment 2, we conducted a third experiment that focused on toddlers' evaluations of social agents with false beliefs. Using the characters, objects, and remote testing methods of Experiment 2, we introduced a second change in the locations of the toys within the boxes, returning each toy to the box that had originally contained it. To compensate for the added length of the study, produced by the second switch in the boxes' locations, we shortened the familiarization events. During familiarization, the bear acted alone, with no aid from either rabbit, and grasped a particular toy in a particular open box while the rabbits observed. Because the boxes were already opened, the bear did not have to struggle to open a box as in Experiments 1 and 2.

Following familiarization, the bear left the scene while the rabbits remained, and two hands moved each toy to the other box and left the stage. Thus, both rabbits observed that the desired toy was now in a different box. Next, the rabbits left the scene and the

hands returned. With no characters onstage, the hands moved the toys back to their original locations and closed the lids on the boxes, inducing false beliefs in the rabbits concerning the desired toy's location.

On each of the final events, the bear appeared at the center of the display, as in Experiments 1 and 2, but now accompanied by a single rabbit, located behind the bear. From this central position, in alternating events, the rabbit with helpful intentions moved to and opened the box where both rabbits had last seen the bear's desired toy, and the other rabbit moved to and opened the box that the bear had approached during familiarization, where both rabbits had last seen the other toy. Because the rabbits began in this central position, this experiment presented the toddlers with evidence that the rabbits chose where to act and which box to act upon.

If the toddlers in Experiments 1 and 2 were unable to track the two rabbits' intentions when they acted under false beliefs, and favored the rabbit whose action was more similar to the previous action of the bear, then the toddlers in Experiment 3 should favor the rabbit that approached and opened the box that the bear had approached during familiarization. By contrast, if toddlers are sensitive to social agents' false beliefs and evaluate social agents based on their intentions, the toddlers should exhibit the opposite preference and favor the rabbit that opened the box where it had last seen the bear's desired toy.

Methods

Participants. Twenty-four full-term 15-month-old toddlers (mean age = 14.90 months; range = 14;10-15;20; 15 female, 9 male) contributed data. One additional participant was excluded due to equipment failure.

Displays and Procedure. Experiment 3's procedure was like that of Experiment 2's False Belief Condition, except as follows. During familiarization, the boxes were open, and the bear was able to grasp the toy inside without help from either rabbit. Because the boxes were already open, there was no pre-familiarization event like that of Experiment 2.

After familiarization, the toys' locations changed twice in the bear's absence: first as the rabbits were present to observe, and again after the rabbits had left the stage. The second change restored the toys to their original boxes, inducing in the rabbits a false belief concerning the desired toy's location.

During the final events, the bear returned at the stage's center, accompanied by one of the two rabbits; each rabbit stood behind the bear on alternating trials. In alternating events, each rabbit opened a different box, giving the bear access to the toy inside, and then the action paused for two seconds before the video looped three additional times.

The reliability of coding in the preference test and in the final events was assessed, as in Experiment 2. For the preference test, the correlations between the two coders were 0.94 and 0.96 for left- and right-looking, respectively. For the final events, the correlation between the two coders' looking times was 0.93.

Results

In Experiment 3's social preference test, the toddlers looked longer to the rabbit that produced the neutral outcome, guided by helpful intentions ($\text{mean}_{\text{neutral-outcome, helpful-intention \%}}$ = 58.4%, 95% CI [53.3%, 63.4%], *SD* = 11.9%, one-sample $t(24) = 3.43$, p =

.002, $d = 0.70$) (Fig. 3B). In contrast, the toddlers again looked equally at the final events in which rabbits acted on different boxes (see SI).

Discussion

In Experiment 3, the toddlers tracked the beliefs of two social agents over two changes in the location of a protagonist's desired toy, and they evaluated the agents based on their intentions. These findings cannot be explained by different levels of interest to final events involving the rabbit with helpful intentions or by imitation-based preferences, and they extend the evidence for belief-based inferences to a situation in which an agent's false beliefs influence its choices of where to act and which object to act upon.

Although the findings from Experiments 1-3 are consistent with the possibility that toddlers incorporate others' beliefs when engaging in social evaluation, the situations that we presented to toddlers were complex, involving changes in objects' location and multiple agents, each with their own mental states. There has been debate about whether toddlers understand false-belief events in the ways that some developmental scientists have proposed they do (Heyes, 2014; Low & Edwards, 2018; Poulin-Dubois et al., 2018). To validate our explanations about the kinds of reasoning that our tasks tap into, we next presented older, verbal children with the events from our tasks, and asked children to describe the events and evaluate the actors.

Experiment 4

In Experiment 4, we presented 6- and 7-year-old children with a version of the puppet show events that were presented to toddlers. We chose to study children at this age because, relative to 4- and 5-year-old children, older children more robustly privilege intentions over outcomes in their evaluations (see Cushman et al., 2013). The present

experiment investigated whether 6- and 7-year-old children represent the beliefs of agents in the events, and whether these children would use those belief representations to infer the intentions of the agents and form intention-based evaluations.

Methods

Participants. Fifty-five 6- and 7-year-old children (mean age = 7.04 years; range = 6.00-7.95; 29 female, 26 male) were tested.

Six additional participants began the experiment but were excluded, based on preregistered criteria, due to inattentiveness (n = 3), equipment failure (n = 2), and procedural error (n = 2), as judged by experimenters who were unaware of the events that children saw, the condition to which a child was assigned, and the role played by each puppet. Of the 55 participants who were not entirely excluded, 13 only contributed partial data due to interference from a child's siblings or caregivers (n = 7), a child reporting that they did not understand a question (n = 5), or equipment failure (n = 1).

Displays. Each child viewed 1 pre-familiarization event, 2 familiarization events, 1 event in which the toys switched locations, and 2 final events, for a total of 6 events. The events were exactly like those of Experiment 2, except that the final events began with the rabbits behind the bear, as in Experiment 3. Thus, the rabbits clearly chose which box to open in the final events. Because an experimenter was able to ask children about their understanding of the events, the events were not looped.

Procedure. In the pre-familiarization event, the experimenter introduced the children to the two rabbits ("bunnies"), the bear, and the opaque blue and green boxes, each with a toy inside matching its color.

After the pre-familiarization event, the children saw 2 familiarization events, like those of Experiment 2. After each familiarization event, the experimenter asked the children to describe the actions of the bear and the rabbit.

After familiarization, the children saw two hands switch the toys' locations, as the bear was off stage, and as the rabbits were present (True Belief Condition) or absent (False Belief Condition). The experimenter asked the children to describe what had happened, and then asked the children to predict where the rabbits would first look for the bear's desired toy when they later came back. This question is comparable to that of verbal false-belief tests for children (e.g., Baron-Cohen et al., 1985).

In the final events, the children saw the bear in front of one of the rabbits, both in the center of the stage, with both boxes closed. In one event, one rabbit opened the original box that previously contained the desired toy, and in the other event, the other rabbit opened the box that now contained the desired toy. After a rabbit had opened a box in an event, the experimenter asked the children to describe what the rabbit had done and to explain why it took that action. When talking about the rabbits, the experimenter referred to the rabbits by their color (e.g., "the pink bunny").

Finally, the experimenter presented the children with the two rabbits, side by side, and asked four questions: after the toys had been switched, which rabbit had wanted to help the bear get its desired toy; after the toys had been switched, which rabbit had wanted to be nicer; based on the rabbits' interactions with the bear, which rabbit the children liked more; and which rabbit the children thought that the bear would like more. We will refer to these four questions as the evaluation questions. We calculated the proportion of answers on which each child more positively evaluated the rabbit with

helpful intentions over the rabbit with less helpful intentions (see SI for analyses of individual questions).

Results

Belief representation. We first examined whether the children tracked the rabbits' beliefs about the location of the bear's desired toy (Fig. 4A). After the toys' locations had changed, when asked where the rabbits would look for the bear's desired toy, 23/26 the children in the True Belief Condition answered that the rabbits would look in the box that currently contained the desired toy (binomial $p < .001$, relative risk = 0.23). By contrast, 28/29 children in the False Belief Condition instead answered that the rabbits would look in the box that used to contain the desired toy (binomial $p < .001$, relative risk = 1.93). Answers differed significantly between conditions ($\chi^2(1) = 36.90$, $p < .001$, Wald's odds ratio = 214.66).

Evaluations. Next, we asked whether the children differently evaluated the rabbits, depending on whether they held true or false beliefs. Within each condition, we examined whether the children chose the rabbit with helpful intentions above chance (50%) when answering the evaluation questions (Fig. 4B). In the True Belief Condition, the children selectively chose the rabbit with helpful intentions, who produced the positive outcome ($\text{mean}_{\text{proportion}} = 73.08\%$, $SD = 31.6\%$, one-sample $t(25) = 3.71$, $p < .001$, $d = 0.72$). In the False Belief Condition, the children selectively chose the rabbit with helpful intentions, who produced the neutral outcome ($\text{mean}_{\text{proportion}} = 64.74\%$, $SD = 19.19\%$, one-sample $t(25) = 3.91$, $p < .001$, $d = 0.76$). The children's choice of rabbit differed significantly between conditions based on outcome (two-sample $t(41) = 5.21$, $p < .001$, $d = 1.44$).

Discussion

In Experiment 4, 6- and 7-year-old children: attributed true and false beliefs to the rabbits based on whether the rabbits had seen the toys' locations changing; inferred the intentions of the agents; and engaged in intention-based evaluations. These verbal evaluations aligned with the toddlers' nonverbal preferences in Experiments 1-3. These findings validate the puppet shows that we had used for toddlers, as children viewing the same displays interpreted them in ways that are consistent with the toddlers' behavior. These findings provide further evidence that these puppet shows elicit reasoning about beliefs.

General Discussion

In three experiments, 15-month-old toddlers engaged in intention-based evaluations of social agents who acted on true and false beliefs. Toddlers' preference for the agent who produced the neutral outcome, acting on outdated information (in the False Belief Conditions), was comparable in strength to their preference for the agent who caused the positive outcome, guided by full information (in the True Belief conditions of Experiments 1 and 2). Across four experiments, we ruled out several lower-level explanations for these findings (e.g., difficulty with goal understanding, imitation-based preferences), and we found that 6- and 7-year-old children's verbal responses to the same agents and events aligned with the nonverbal responses of toddlers. Thus, toddlers' social evaluations privilege intentions over outcomes, and they are sensitive to the mental representations on which agents' actions are based.

These findings in toddlers are striking, given young children's well-documented struggles to see past outcomes in verbal tasks assessing their social evaluations (Cushman

et al., 2013; Piaget, 1965; Yuill & Perner, 1988; Zelazo et al., 1996) and given the difficulties that even adults can face when presented with agents who hold false beliefs (Apperly et al., 2008). The present findings contribute to a growing body of evidence that infants' and young children's social evaluations are sensitive to others' intentions (Hamlin, 2013; Hamlin et al., 2013; Kanakogi et al., 2017; Margoni & Surian, 2020; Martin et al., 2021; Woo et al., 2017), and they provide evidence that toddlers evaluate agents in accord with their beliefs and intentions, rather than the outcomes that their actions produce.

The consistency of the present findings contrasts with the inconsistent evidence for early sensitivity to agents' beliefs in minimally social contexts. Infants and toddlers do not consistently predict an agent's actions from its beliefs when presented with enactments of classic false belief scenarios, in which agents act for their own benefit (Phillips et al., 2020; Poulin-Dubois et al., 2018). Yet, human life is centered around cooperation: Especially in early childhood, humans often depend on others to accomplish goals and learn new skills and knowledge (Gweon, 2021; Hrdy, 2011; Tomasello & Carpenter, 2007). We suggest that infants and toddlers reason about agents' beliefs and intentions more readily when the agents' actions have social consequences. The beliefs of a prosocial agent support inferences about its intentions, which may reveal social qualities such as cooperativeness or generosity. These inferences have implications not only for our evaluations but for our own social decisions: Is this individual potentially a good or bad social partner for me? If I engage with this individual, are good or bad consequences likely to follow? This proposal is consistent with evidence that in verbal tasks, young children are more sensitive to others' beliefs in more strongly social

contexts (Asaba & Gweon, 2019; Tsoi et al., 2020, 2021; Wellman et al., 2001). To our knowledge, no experiments test this possibility directly in infants or younger toddlers.

The present findings raise questions concerning the nature of early mental state attributions. In these experiments, as in most past research probing false-belief understanding in toddlers (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017) and nonhuman primates (Krupenye et al., 2016), inferences about agents' mental states depended on the events that agents did and did not observe. Adults, however, make finer distinctions among mental states, based on agents' differing experiences of the same perceptually accessible objects (Surtees et al., 2012). Different people may observe the same object but see different things: A single menu, for example, may be readable to one person at a table but not to someone who faces them. Relative to research probing an understanding of false beliefs, less research, to our knowledge, has probed the nonverbal capacities of human infants, toddlers, or nonhuman animals to reason about the diverse mental states that the same observable objects can elicit in different observers (F. Buttelmann et al., 2015; Karg et al., 2016; Luo & Beck, 2010; Moll & Meltzoff, 2011). We look forward to research that further probes such capacities.

In conclusion, the present experiments reveal that early social evaluation takes account of the beliefs of social agents. Toddlers view others as having intentions that are modulated by their representations of the world, and those intentions bear on the social value of their actions. Because mental states offer a window into the qualities of mind that are predictive of a social partner's future actions, desires, and commitments, an early-emerging focus on social agents' mental states may foster children's learning to navigate their social world.

References

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, *106*(3), 1093–1108. https://doi.org/10.1016/j.cognition.2007.05.005

Asaba, M., & Gweon, H. (2019). *Young children can rationally revise and maintain what others think of them*. PsyArXiv. https://doi.org/10.31234/osf.io/yxhv5

Astington, J. W. (2004). Bridging the gap between theory of mind and moral reasoning. *New Directions for Child and Adolescent Development*, *2004*(103), 63–72.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

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. https://doi.org/10.1016/j.cognition.2009.05.006

Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology*, *131*, 94–103.

Casstevens, R. M. (2007). jHab: Java habituation software (version 1.0. 2)[computer software]. *Chevy Chase, MD*.

Colomer, M., Bas, J., & Sebastian-Galles, N. (2020). Efficiency as a principle for social preferences in infancy. *Journal of Experimental Child Psychology*, *194*, 104823.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Geraci, A., Simion, F., & Surian, L. (2022). Infants' intention-based evaluations of distributive actions. *Journal of Experimental Child Psychology*, *220*, 105429. https://doi.org/10.1016/j.jecp.2022.105429

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*.

Hamlin, J. K. (2013a). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193. https://doi.org/10.1177/0963721412470687

Hamlin, J. K. (2013b). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451–474. https://doi.org/10.1016/j.cognition.2013.04.004

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me= bad: Infants prefer those who harm dissimilar others. *Psychological Science*, *24*(4), 589–594.

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. https://doi.org/10.1111/desc.12017

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559. https://doi.org/10.1038/nature06288

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. https://doi.org/10.1111/j.1467-7687.2010.00951.x

Henry, O. (2012). *The Gift of the Magi and other short stories*. Courier Corporation.

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, 567–582.

Holland, C., & Phillips, J. S. (2020). *A theoretically driven meta-analysis of implicit theory of mind studies: The role of factivity*. 1749–1755.

Hrdy, S. B. (2011). *Mothers and others*. Harvard University Press.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224–234. https://doi.org/10.1016/j.cognition.2008.08.010

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.

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2016). Differing views: Can chimpanzees do Level 2 perspective-taking? *Animal Cognition*, *19*(3), 555–564.

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. https://doi.org/10.1016/j.cognition.2011.01.006

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*(8), 357–359.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Low, J., & Edwards, K. (2018). The curious case of adults' interpretations of violation-of-expectation false belief scenarios. *Cognitive Development*, *46*, 86–96.

Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent. *Developmental Science*, *14*(2), 453–460.

Luo, Y., & Beck, W. (2010). Do you see what I see? Infants' reasoning about others' incomplete perceptions. *Developmental Science*, *13*(1), 134–142.

Margoni, F., & Surian, L. (2016). Explaining the U-shaped development of intent-based moral judgments. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00219

Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology*, *54*(8), 1445–1455.

Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, *14*(6), 1406–1416. https://doi.org/10.1111/j.1467-7687.2011.01085.x

Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before Belief. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20000618

Piaget, J. (1965). *The moral judgment of the child.* Routledge & K. Paul.

Pinto, J. (1996). *XHAB64*.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & 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. https://doi.org/10.1016/j.cogdev.2018.09.005

Powell, L. J. (2021). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*.

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004

Powell, L. J., & Spelke, E. S. (2018). Third-party preferences for imitators in preverbal infants. *Open Mind*, *2*(2), 61–71.

Salvadori, E., Blazsekova, T., Volein, A., Karap, Z., Tatone, D., Mascaro, O., & Csibra, G. (2015). Probing the strength of infants' preference for helpers over hinderers: Two replication attempts of Hamlin and Wynn (2011). *PloS One*, *10*(11), e0140570.

Schlingloff, L., Csibra, G., & Tatone, D. (2020). Do 15-month-old infants prefer helpers? A replication of Hamlin et al.(2007). *Royal Society Open Science*, *7*(4), 191795.

Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Smetana, J. G., Jambon, M., & Ball, C. (2014). The social domain approach to children's moral and social judgments. *Handbook of Moral Development*, *2*, 23–45.

Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022). Online measures of looking and learning in infancy. *Infancy*, *27*(1), 4–24.

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's theory of mind become explicit? A review of longitudinal findings. *Child Development Perspectives*, *14*(3), 171–177.

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition*, *95*(1), 1–30.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Surtees, A. D. R., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology*, *30*(1), 75–86. https://doi.org/10.1111/j.2044-835X.2011.02063.x

Thomas, A. J., & Sarnecka, B. W. (2019). Infants choose those who defer in conflicts. *Current Biology*, *29*(13), 2183–2189.

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

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125.

Tsoi, L., Hamlin, J. K., Waytz, A., Baron, A. S., & Young, L. L. (2021). A Cooperation Advantage for Theory of Mind in Children and Adults. *Social Cognition*, *39*(1), 19–40. https://doi.org/10.1521/soco.2021.39.1.19

Tsoi, L., Hamlin, J. K., Waytz, A., Baron, A. S., Young, L., & Tsoi, L. (2020). *False belief understanding for negative versus positive interactions in children and adults*.

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

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

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. https://doi.org/10.1016/0010-0277(83)90004-5

Woo, B. M., Liu, S., & Spelke, E. (2021). Open-minded, not naïve: Three-month-old infants encode objects as the goals of other people's reaches. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Woo, B. M., & Spelke, E. (2022). *Infants and toddlers leverage their understanding of action goals to evaluate agents who help others.*

Woo, B. M., & Spelke, E. S. (2020). How to help best: Infants' changing understanding of multistep actions informs their evaluations of helping. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 384–390.

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. https://doi.org/10.1016/j.cognition.2017.06.029

Woo, B. M., Tan, E., & Hamlin, K. (in press). Human morality is based on an early-emerging moral core. In *Annual Review of Developmental Psychology*. https://doi.org/10.31234/osf.io/98d36

Woodward, A. L. (2009). Infants' grasp of others' intentions. *Current Directions in Psychological Science*, *18*(1), 53–57. https://doi.org/10.1111/j.1467-8721.2009.01605.x

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, *11*(1), 73–77.

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. https://doi.org/10.1073/pnas.0701408104

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, *24*(3), 358–365.

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, *67*(5), 2478–2492.

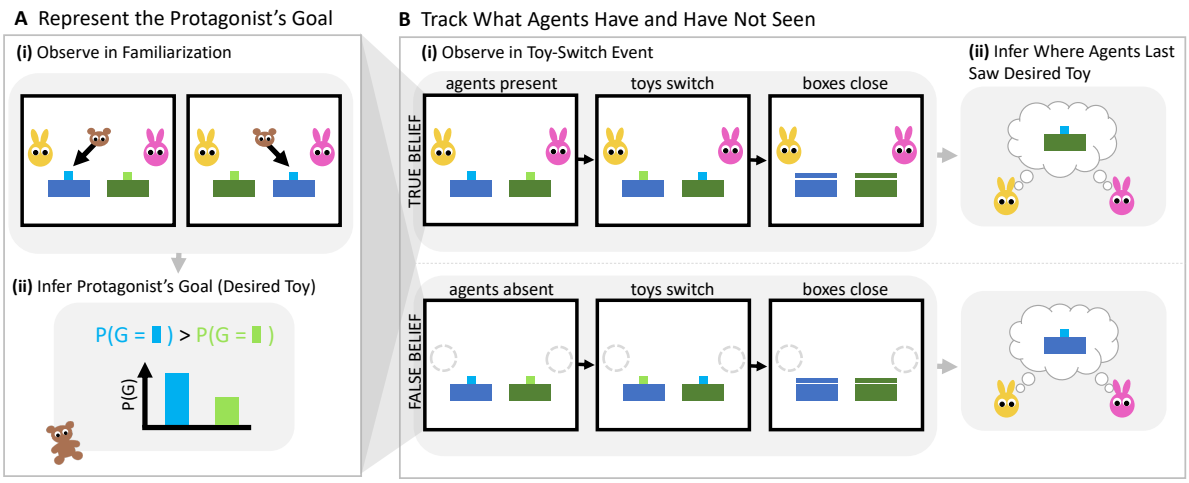

Figure 1. A conceptual schematic of the inferences about the protagonist's goal (A) and the agents' representations of the contents of the boxes (B) that Experiments 1 and 2 assess. In familiarization, toddlers observe the protagonist (the brown bear) repeatedly acting on the toy within one of the two boxes (in (A i), the toy in blue), as two other agents (rabbits clothed in pink and yellow) observe. From observing this preferential behavior, toddlers are expected to infer that the protagonist's goal is to obtain one toy (here, the blue one) (A, ii), and that the two rabbits have knowledge of this goal. In the toy-switch event, toddlers observe that the rabbits are present in the True Belief Condition and absent in the False Belief Condition, as the toys in the two boxes exchange locations (B, i). Toddlers are challenged to infer (B, ii) that the agents have an accurate representation of the desired toy's location in the True Belief Condition, and an outdated

representation of the desired toy's location in the False Belief Condition, based on where agents had last seen the desired toy.

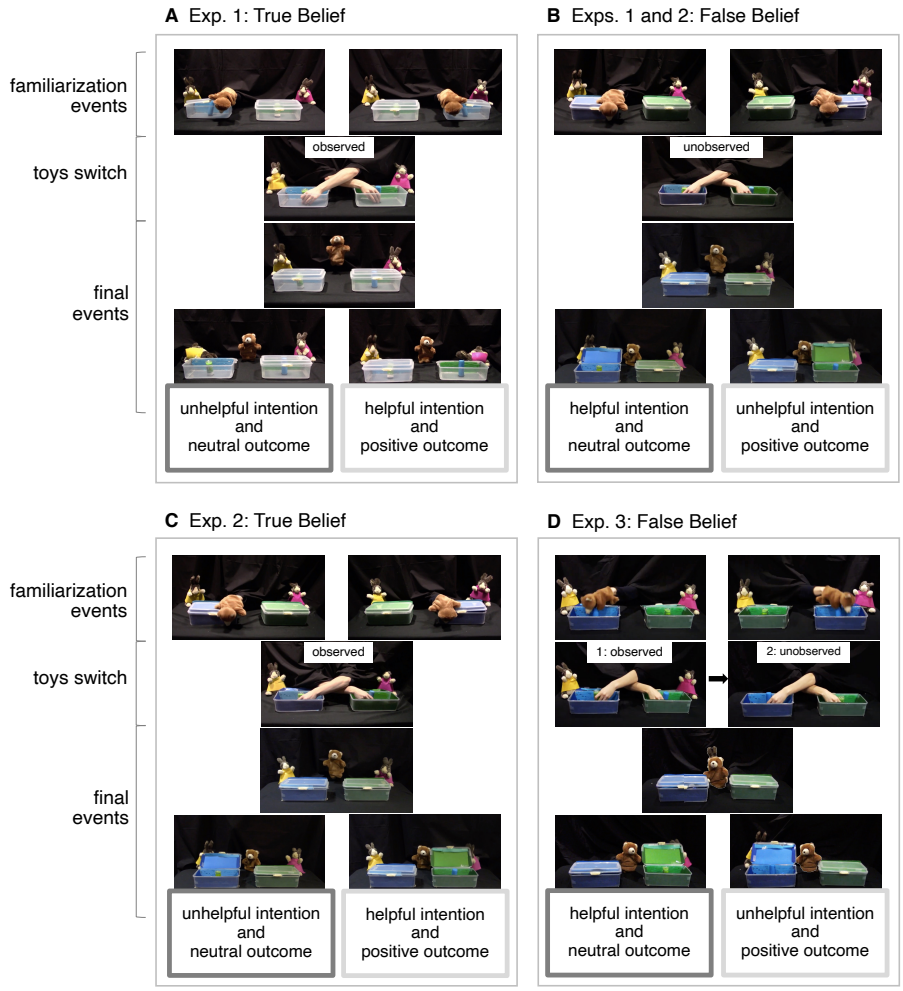

Figure 2. Events presented to toddlers in Experiments 1 (A, B), 2 (B, C), and 3 (D). In familiarization events, the bear retrieved one toy from the same box, observed by two rabbits. In toy-switch events, a pair of hands switched the boxes' contents, either as the rabbits were present or absent to observe. In the final events, each rabbit opened a different box. In the False Belief Conditions (B, D), the rabbits did so in a manner that

left their contents unseen by the rabbits, maintaining their ignorance of the movement of the desired toy.

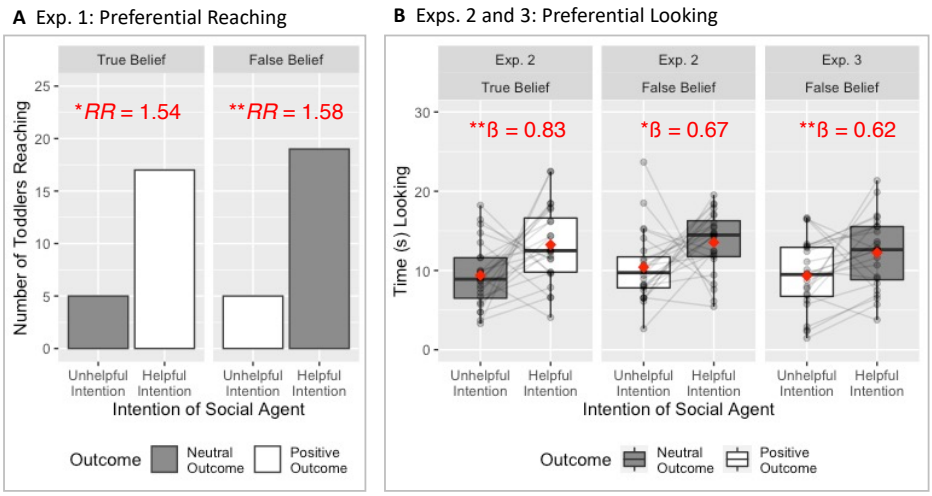

Figure 3. Results in Experiment 1 (A) and Experiments 2 and 3 (B). Panel A depicts the number of toddlers reaching for each social agent by condition in the social reaching preference test; *RR* indicates relative risk. Panel B depicts the mean time each toddler looked to each rabbit by condition in 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 (*$p < .05$, **$p < .01$).

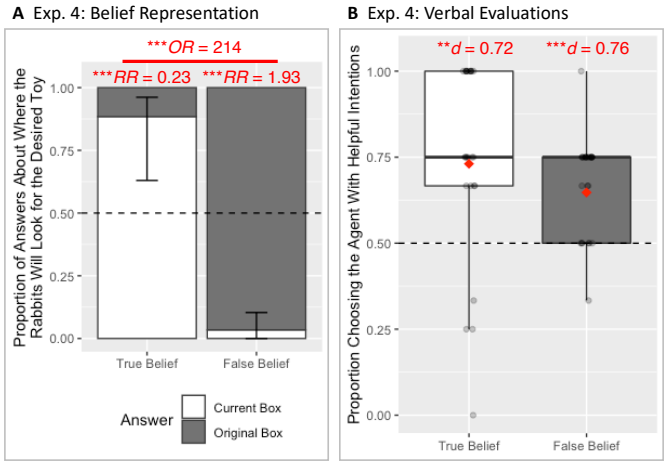

Figure 4. Results in Experiment 4. Panel A depicts children's answers when asked where the rabbits would look for the bear's desired toy, in relation to the rabbits' beliefs about the toy's location. *RR* indicates relative risk, *OR* indicates odds ratio, and error bars indicate bootstrapped 95% confidence intervals. Panel B depicts the proportion of children who more positively evaluated the agent with helpful intentions within each condition. Red diamonds indicate means, dots indicate data from individual children, and Cohen's *d* indicates standardized effect size. Across panels, asterisks indicate significant differences (**$p < .01$, ***$p < .001$).