scientific reports

Check for updates

OPEN

Children's attribution of mental states to humans and social robots assessed with the Theory of Mind Scale

Elizabeth J. Goldman[1]✉, Anna-Elisabeth Baumann[2], Laetitia Pare[3], Jenna Beaudoin[3] & Diane Poulin-Dubois[3]

The present work examined children's attribution of psychological properties to inanimate agents in two experiments. In Study 1, an Interview Task and the Theory of Mind Scale (ToM Scale) were administered to 4-year-olds with either a human or a humanoid robot (NAO) protagonist. Parents also completed the Children's Social Understanding Scale (CSUS) to assess children's Theory of Mind skills. Overall, children performed similarly on the Interview and the ToM Scale. Theory of Mind skills (CSUS) did not predict performance on either task (ToM Scale or Interview). In Study 2, 5-year-olds were tested with figurines of different humanoid robots for the ToM Scale. Additionally, a Property Projection Task assessed biological, psychological, sensory, and artifact attributions to people, robots, animals, and artifacts. The results indicated that children attributed mental states similarly to the robots and the humans in the ToM Scale but did not anthropomorphize the robots in the Property Projection Task. In contrast to Study 1, the parental measure of children's ToM skills (CSUS) predicted performance on the ToM Scale in the Human Condition. Overall, the present findings indicate that mentalizing is generalized to humanoid robots by preschoolers, particularly when child-friendly scenarios are used.

Keywords Theory of Mind, Inanimate agents, Robots, Interview, Anthropomorphism, Animacy

Anthropomorphism is the attribution of human properties to non-human entities[1]. In a first wave of developmental studies on anthropomorphism, children were often tested about the biological properties (e.g., grow, reproduce) of animals, artifacts, and plants[2–8]. Much of this work was conducted through interviews where children answered questions while presented with images of animals, artifacts, and plants.

Anthropomorphism of robots

Similarly to plants, researchers have tested children using another ambiguous item, robots[9–15]. Robots offer a unique way to test anthropomorphism, as robots are artifacts that can be designed to exhibit animate characteristics. Recent studies have used experimental tasks and detailed, structured questionnaires to test developmental changes in anthropomorphism. In two recent studies, children aged 3 and 5 attributed biological insides to animals and mechanical ones to artifacts. Notably, the younger children were unsure about the insides of robots, but the 5-year-olds knew that humanoid and non-humanoid robots had mechanical insides[11,16].

In addition to investigating the biological properties of animates and inanimates, researchers have also studied children's anthropomorphism of mental properties (thinking, feeling)[9,11,12,17,18]. For example, Saylor et al.[15] found that 3- and 4-year-olds attributed properties of living things to an image of a girl more than to a camera or robot. However, only the 4-year-olds categorized the robot as a machine. These findings indicate that by 4 years, children understand that robots are machines that are different from people. A recent review by Goldman and Poulin-Dubois[19] concludes that children anthropomorphize less as they age, but social robots appear to be an exception, with anthropomorphizing of robots occurring throughout childhood.

To better assess anthropomorphism, a team of Italian researchers has developed the standardized and widely used Attribution of Mental States Questionnaire (AMS-Q) to evaluate the attribution of sensory and mental states to various human and non-human agents[9,12,18]. For example, in a recent study with the AMS-Q, Manzi et

[1]Children and Technology Laboratory, Department of Psychology, Yeshiva University-Stern College for Women, New York, NY, USA. [2]Language and Cognitive Development Laboratory, Department of Psychology, University of Calgary, Calgary, Canada. [3]Cognitive and Language Development Laboratory, Department of Psychology, Concordia University, Montréal, Canada. ✉email: elizabeth.goldman@yu.edu

al.[12] asked 5-, 7-, and 9-year-olds about the mental states of two different robots (Robovie and NAO), which were depicted via visual images. The youngest children in the study, the 5-year-olds, attributed more mental states to both robots than the older age groups.

Jipson et al.[17] also tested anthropomorphism by administering a questionnaire, the Property Projection Task, to 3- and 5-year-olds. Within the Property Projection Task, children responded to a series of interview questions about a robotic dog, a rodent, and a toy car to assess their understanding of these items across various domains (biological, psychological, sensory, artifact). Regardless of age, children tended to attribute biological properties more to the rodent than the robotic dog or toy car.

Beyond interviews

Although interviews offer an easy way to test anthropomorphism in children, they bring methodological limitations because they cannot be used with young children. Additionally, interviews with young children may yield a "yes bias." Given these limitations, recent work has examined children's ability to anthropomorphize with experimental tasks that provide context and better reflect children's mentalizing in everyday life, increasing ecological validity. A review by van Straten et al.[20] found that younger children anthropomorphized robots more than older children.

In developmental science, the gold standard for measuring children's ability to attribute mental states to others is the Theroy of Mind Scale (ToM Scale) developed by Wellman and Liu (2004)[21–25]. It consists of five vignettes that feature a human protagonist who is experiencing mental states ranging from simple desires to hidden emotions. To our knowledge, only one study has used the ToM Scale to assess anthropomorphism in robots. Zhang and colleagues[26] administered a change of location false belief task and an unexpected contents task, modified from the ToM Scale, which featured a humanoid robot (NAO) as the protagonist. Researchers found that children aged 5 to 7 attributed false beliefs to the robot.

The studies presented here aimed to examine whether 4- and 5-year-olds will attribute mental states to humanoid robots and whether this is comparable to how children attribute mental states to humans. Both studies were conducted over Zoom. Study 1 examined whether children attribute mental states to humanoid robots. Children responded to a series of interview questions; some were modified and adapted from the AMS-Q[12]. Children also completed the ToM Scale[27] using human or robot figurines to illustrate the protagonist featured in the vignettes. To our knowledge, this is the first study to administer the complete ToM Scale with non-human protagonists. Finally, parents completed the Children's Social Understanding Scale (CSUS)[28], which assessed young children's Theory of Mind. As preschoolers have been shown to anthropomorphize, we predicted that children in both conditions would attribute mental states to the robot and human agent equally. We did not hypothesize that children in the human condition would perform differently than children assigned to the humanoid robot condition on the individual items of the ToM scale. As direct tasks provide more context, we anticipated that children would anthropomorphize more on the ToM Scale than the Interview Task.

A follow-up study was conducted to address some limitations of Study 1 and to examine mental state attribution in slightly older children. In Study 2, 5-year-olds were administered the ToM Scale, which featured humanoid robot figurines that varied in morphology. To better understand children's willingness to anthropomorphize and attribute mental states, children were asked about various animate and inanimate items via an adapted version of Jipson et al.'s[17] Property Projection Task. The CSUS was also administered in Study 2. As in Study 1, we expected children to attribute mental states equally to the human and robot protagonists on the ToM Scale. We did not predict differences between the human and humanoid robot groups on the items of the ToM scale. Based on prior work, we hypothesized that children would judge animacy accurately for items they were familiar with (e.g., humans, animals, and artifacts) but be uncertain about the animacy of robots when responding to interview questions.

Study 1

Method

Participants

Participants included 102 children who were four years of age (*Mage* = 54 months, 16 days; *Nmale* = 52). The children were recruited from a university participant pool and through social media advertisements. The sample consisted mainly of children of Caucasian (42%) or Asian (35%) descent, and the remainder of the sample was of mixed ethnicity (20%) or did not provide their ethnicity (3%). A majority of children in the sample were from high (more than $100,000; $n = 64$) or middle (more than $50,000 but less than $100,000; $n = 22$) socioeconomic status (SES) families. The remainder of the sample were from low-SES (family income less than $50,000) families ($n = 9$) or opted not to report family SES on the demographic form ($n = 7$). The sample consisted of Canadian and American children tested in French ($n = 8$) or English ($n = 94$).

Power was calculated for the General Linear Model using G*Power 3.1. The required power was estimated using the linear multiple regression: fixed model, R^2 deviation from zero analysis. Effect size was estimated at medium 0.15, alpha error probability was entered at 0.05, power at 0.80, and the number of predictors entered was 5 (Condition, Age, SES, CSUS, and Robot Exposure). For this analysis, a total sample size of 92 is required. For the mixed-effects model (MEM), power was calculated the same way, adding in an additional variable measuring variation between the different interview subscales for a total of 6 predictors. Therefore, the required sample size for our MEM is 98. Therefore, our sample size is sufficient for our analyses.

An additional four participants were tested but excluded due to experimenter error ($n = 1$), parental interference ($n = 1$), and distractedness/failure to complete the study ($n = 2$). We examined the Robot Exposure Questionnaire to determine whether children had regular exposure to or interactions with robots. The questionnaire found that 16% of parents responded that their children watched a television show or movie that features robots, and 12% of parents reported having a robot at home, but only 5% said their child had regular interactions

with a robot, but none of the robots were similar to the one used in the study. Therefore, no participants were excluded due to their familiarity with robots. The study was approved by the University's Human Research Ethics Committee (certificate of ethical acceptability #10000548) and was conducted in accordance with the guidelines and regulations outlined by the Ethics Committee.

Measures

Robot Exposure Questionnaire

Parents completed a short questionnaire about their child's exposure to and experiences with robots. Questions included (1) whether the family possessed a robot at home, (2) whether the child plays with a robot at home, (3) whether the child interacts with a robot frequently, and (4) how often children play video games or watch TV/movies that feature robot characters. The questions were all yes/no/unsure forced response.

Wellman & Liu Scale

To measure Theory of Mind skills, the well-validated ToM Scale[27] was administered. It consists of five items of increasing difficulty (Diverse Desires, Diverse Beliefs, Knowledge Access, Contents False Belief, and Hidden Emotion). Figurines (Human and Robot) and props are used to illustrate the story and characters in each of the tasks; see Table 1 and Fig. 1.

Following the standard procedure, the five items of the ToM Scale were always administered in the same order (Diverse Desires, Diverse Beliefs, Knowledge Access, Hidden Contents False Belief, Hidden Emotions), with the items being presented in increasing difficulty, as described in the original study[27]. All robot and human figurines were assigned a name (e.g., Dash, Sam). For a full list of proper names used, see Appendix A.

Interview Task

The other task was an interview consisting of 14 questions (see Appendix B for a full list of questions). Some questions were modified from Manzi and colleagues'[12] Attribution of Mental States Questionnaire (AMS-Q), a validated measure originally administered in Italian. For the present study, the interview was conducted in either English or French as these were the official languages of the population tested. Two subsets of items were administered from the AMS-Q: (1) Epistemic (e.g., "Do you think this robot/person can learn?") and (2) Intentions and Desires (e.g., "Do you think this robot/person can make a wish?"). A third subset was created by the experimenters (False Beliefs, e.g., "Can this robot/person believe something that is incorrect?"). The subscales used and created were selected because they aligned with the constructs measured on the Wellman and Liu Scale. There were five questions for the Epistemic subset, five for the Intentions and Desires subset, and four for the False Beliefs subset. Depending upon the assigned condition, children were presented with a picture of either a human-looking robot (NAO) or a participant gender-matched adult human and were asked the

Task	Description	Image
Diverse Desires (DD)	The child is asked if they prefer to eat a cookie or a carrot. The child is then told that the agent (robot or human) prefers to eat the opposite. The child is asked what the agent will choose.	
Diverse Beliefs (DB)	The child is asked whether they think the agent's (robot or human) cat is hiding in the bushes or the garage. The child is told the agent believes the cat is hiding in the opposite location. The child is asked where the agent will look for the cat.	
Knowledge Access (KA)	The child sees what's inside a box. The child is told the agent (robot or human) has never seen inside the box. The child is asked whether the agent knows what is inside the box. The child is also asked if the agent has seen inside the box.	
Contents False Belief (FB)	The child is shown a box of crayons and is asked what they believe is inside. The researcher opens the box and reveals a toy pig inside the box of crayons. The agent (robot or human) is introduced, and the child is told that the agent has never seen inside the box of crayons. The child is asked what the agent will think is inside the box of crayons and whether the agent had seen inside the box?	
Hidden Emotion (HE)	The child is asked to judge how the agent (robot or human) feels inside and how the agent actually looks on their face.	

Table 1. Brief descriptions of the items in the Theory of Mind Scale.

Fig. 1. The robot and human figurines used in the ToM Scale Task in Study 1.

Fig. 2. The robot and human stimuli used in the interview in Study 1.

interview questions about the robot or the human shown in the picture, see Fig. 2. The questions were a forced choice format so that children could respond "yes" or "no" to each interview question. If the child responded, "maybe," or failed to provide a clear answer, they were re-prompted by the experimenter and asked to make their best guess. The order in which the three subsets of interview questions (Epistemic, Intentions and Desires, and False Belief) were asked was randomized across the participants.

Children's Social Understanding Scale (CSUS)

Parents completed the Children's Social Understanding Scale (CSUS), a parent-report measure of ToM developed by Tahiroglu et al.[28]. This parental report measure was used to measure a child's ToM in a non-experimental context. Additionally, it allowed for the examination of whether parental perceptions of their child's ToM aligned with children's ToM skills, as tested experimentally. The questionnaire is comprised of six subscales: Beliefs (e.g., beliefs can differ about the same situation); Knowledge (e.g., people can have different levels of knowledge); Perception (e.g., reality and appearances are not always the same); Desire (e.g., people can desire different things); Intention (e.g., same intentions may have different outcomes); and Emotion (e.g., people may feel differently about the same situation). There are seven items per subscale for a total of 42 statements. Parents rated their child's ability for each item on a Likert scale ranging from 1 *(definitely untrue for my child)* to 4 *(definitely true for my child)*. Parents also had the option to respond "don't know" if they lacked insight into their child's behavior on a particular item. Parents received the questionnaire by email and filled it out before or

after the testing session. Either the long form of the CSUS or the French adaptation, l'Échelle de compréhension sociale des enfants (ÉCSE;[29]), was administered depending on the parent's dominant language.

Materials

Distinct figurines (Human or Robot) were used for each of the five ToM Scale items. The figurines were approximately 5 inches tall and 2 inches wide. All the robot figurines varied in color, were 3D printed, and depicted the robot NAO. In addition to the figurines, printed images (see Table 1) were used as props in the vignettes (e.g., cookie, carrot, garage, bushes). Other materials included a 5 × 5 inch box with a toy dog used for the Knowledge Access item and a Crayola crayons box with a toy pig used for the Contents False Beliefs item. While the interview questions were asked, participants were shown an image of either a man or woman (gender-matched to the participant, Human Condition) or the robot NAO (Robot Condition).

Procedure

Prior to the Zoom session, parents completed a consent form, the CSUS, and a demographic form. Parents also completed a Robot Exposure Questionnaire to gauge whether children had regular interactions with robots or if they frequently watched television or movies that featured robot characters. The Zoom session began with a PowerPoint presentation during which the experimenter introduced the study and requested verbal informed consent from the parent and verbal assent from the child. Before the first task was administered, the experimenter assessed that the participant's environment was free of distractions and confirmed that the child could be seen on camera and heard on the microphone. Parents were requested to use a tablet or a computer with a minimum screen size of 8 inches to join the Zoom session to ensure that the screen size was large enough for the child to clearly see the stimuli.

Participants were randomly assigned to one of two conditions (Human or Robot). In the Human Condition, children were presented with different human figurines or props for each task. In the Robot Condition, the child was presented with figurines and props of a humanoid robot. The order of the two tasks (ToM Scale and Interview) was counterbalanced. After both tasks were completed, parents were debriefed on the study's goals and invited to ask any questions.

Scoring

Children received a passing or failing score for each ToM Scale item. Therefore, scores on this task ranged from 0 to 5. For the interview, the child received a point for each "yes" answer, with the total score ranging from 0 to 14. Each child received a mean total score for the CSUS (out of 4). For the Robot Exposure Questionaire, any data cells where parents had answered "unsure" were left blank. Otherwise, children received scores based on their robot exposure, as reported by the parent. For each question, 'yes' was coded as 1 and 'no' as 0. The score for all 4 questions was added together to create an overall Robot Exposure Score per child, with higher scores indicating more familiarity with and experiences with robots. To categorize Socio-Economic Status as a variable, SES was split into 3 categories, with families making less than $50,000 classified as low SES, $50,0000 to $100,000 classified as mid, and over $100,000 classified as high.

Results and Discussion

Data Analysis

The data for each task (ToM Scale, Interview) and the CSUS was analyzed independently. Then, correlations were run to determine if performance on one task was correlated with the CSUS. For cross-task analyses, raw scores were turned into proportions. Unless otherwise specified, statistical analyses were performed using JASP 0.18.3[30].

Theory of Mind Scale

Children in the Human Condition ($M = 2.92$, $SD = 0.87$; $t(50) = 3.47$, $p = 0.001$, $d = 0.49$) performed as expected for their age on the ToM Scale. Children in the Robot Condition ($M = 2.67$, $SD = 0.95$; $t(50) = 1.25$, $p = 0.22$, $d = 0.18$) had a slightly lower score. However, children assigned to the Human Condition ($n = 51$) performed statistically similarly on the ToM Scale compared to children assigned to the Robot Condition ($n = 51$), as shown by a Generalized Linear Model using the *glm* function run in R Version 4.3.3[31] ($\chi^2(1) = -0.24$, $p = 0.24$, overall $R^2 = 0.02$). A number of factors were entered into the GLM (Age in months, Family SES, Child Robot Exposure, and the CSUS score) but none reached significance. Next, each item of the ToM Scale was analyzed independently to determine the success rate on each item and whether the children in one condition outperformed the other on that item.

As shown in Table 2, binomial tests revealed that children in the Human Condition performed above chance level on the Diverse Desires and the Diverse Beliefs items. Children assigned to the Human Condition were at chance level for the Knowledge Access item and below chance level for the Contents False Belief and the Hidden Emotion items. Regarding the Robot Condition, binomial tests showed that children performed at chance level on the Diverse Desires, Knowledge Access, and the Hidden Emotion items but performed above chance level on the Diverse Beliefs item and below chance level on the Contents False Belief item.

Children in the Human Condition outperformed children in the Robot Condition for the Diverse Desires item ($X^2(1, N = 102) = 16.22$, $p < 0.001$). In contrast, children in the Robot Condition outperformed children in the Human Condition for the Contents False Belief item ($X^2(1, N = 102) = 4.29$, $p = 0.038$). Importantly, children in both conditions performed below chance for the Contents False Belief item; thus, this difference is not considered meaningful. Despite their poor performance on the Contents False Belief item, when asked, "What do you think is inside the box of crayons?" a majority of children, across conditions, correctly responded with crayons (0.86, $p < 0.001$). Thus, children knew what the box should contain but failed to recognize the agent

ToM scale item	Human condition	Robot condition	Difference between the human and robot conditions
Diverse Desires	94.1***	60.8	$X^2(1, N=102)=16.22, p<0.001$***
Diverse Beliefs	82.4***	70.6**	$X^2(1, N=102)=1.96, p=0.16$
Knowledge Access	64.7*	58.8	$X^2(1, N=102)=0.37, p=0.54$
Contents False Belief	15.7***	33.3*	$X^2(1, N=102)=4.29, p=0.038$*
Hidden Emotion	35.3*	43.1	$X^2(1, N=102)=0.66, p=0.42$

Table 2. Percentage of children who passed each item on the Theory of Mind Scale in each condition. * Represents a *p*-value of less than 0.05, ** below 0.01, *** below 0.001 in comparison to chance.

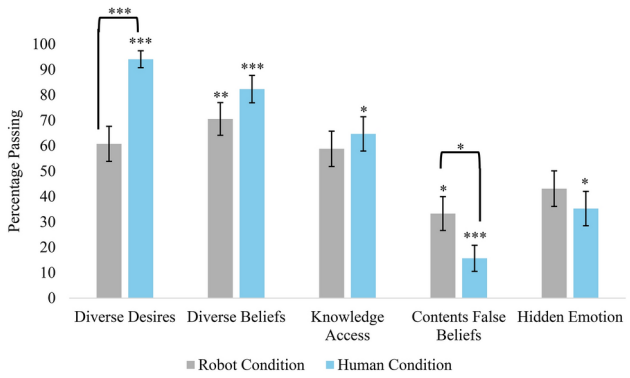

Fig. 3. Mean percentage of children who passed each item of the Theory of Mind Scale in each condition. Error bars represent standard error. * Represents a *p*-value of less than 0.05, ** below 0.01, *** below 0.001. The * directly above the column represent significance against chance. The * in between columns represent significance between groups.

Measure	Human condition		Robot condition		Difference between the human and robot conditions
	M	*SD*	*M*	*SD*	
Epistemic score (out of 5)	3.65***	1.51	2.61	1.74	$t(100)=3.22, p=0.002$**
Intentions and Desires score (out of 5)	3.47***	1.53	3.33***	1.55	$t(100)=0.45, p=0.65$
False Belief score (out of 4)	2.14	1.27	2.20	1.20	$t(100)=-0.24, p=0.81$

Table 3. Mean scores on the Interview (total scores and subsets) in each condition. ** Represents a *p*-value of below 0.01, *** below 0.001 in comparison to chance.

would also expect the box to contain crayons in both conditions. Children's performance did not differ based on condition for the Diverse Beliefs, Knowledge Access, and Hidden Emotions items (see Table 2 and Fig. 3).

Interview Task

Children in the Human Condition ($M=9.26$, $SD=3.24$; $t(50)=4.97$, $p<0.001$, $d=0.70$) and the Robot Condition ($M=8.14$, $SD=3.75$; $t(50)=2.16$, $p=0.035$, $d=0.30$) performed well on the Interview. We ran a linear mixed-effects model, fit by REML, testing the effects of condition on children's Interview performance in Jamovi version 2.3[32]. Interview Question Type (Epistemic, Intentions and Desires, False Beliefs), SES, and Robot Exposure were entered as factors into the model while Age in months and the CSUS score were entered as covariates. The model was clustered by participant. To capture random effects, an intercept for participant was added. Coding for all factors was simple. Condition did not affect children's interview responses, with children performing similarly in the Human and Robot conditions; $b=-1.85$ [−9.85, 6.15], SE = 4.08, $p=0.65$. The overall conditional R^2 was 0.52, with just over half of the variance explained by the model. However, none of the factors or covariates reached significance except for the difference between the False Belief and Epistemic interview questions ($b=-0.93$ [−1.27, -0.60], SE = 0.17, $p_{bonferroni}<0.001$). Post-hoc tests reveal that children performed better for the Epistemic ($t(178)=5.51$, $p_{bonferroni}<0.001$), and Intentions and Desires ($t(178)=-7.35$, $p<0.001$) subsets when compared to the False Belief subset.

When each set of questions was considered, children in the Human Condition performed above chance on both the Epistemic and Intentions and Desires subsets. However, children performed at chance for the False Belief subset. For the Robot Condition, children performed above chance on the Intentions and Desires subset of questions but at chance for the Epistemic and False Belief subsets, see Table 3. When compairing performance

between the groups for each interview subset, we found that children assigned to the Human Condition outperformed children in the Robot Condition on the Epistemic subset. There was no difference in performance on the other subsets; see Table 3.

Children's Social Understanding Scale (CSUS)

Only two parents failed to complete the CSUS for a final sample of 100[28]. Participants' total CSUS score was calculated ($M = 3.18$, $SD = 0.39$). Overall, parents rated their children's Theory of Mind in line with those reported in prior research at a similar age[28,29].

Exploratory Inter-tasks Comparison

With a sample combining all children, the total score on ToM Scale and the total score on the Interview were not correlated ($r(100) = -0.15$, $p = 0.15$). Split by condition, the Human Condition tasks (ToM Scale, Interview) were not significantly correlated ($r(49) = 0.03$, $p = 0.84$). There was a significant negative correlation, however, between the tasks for the Robot Condition ($r(49) = -0.33$, $p = 0.019$), with better performance on the ToM Scale predicting worse performance on the Interview.

Correlations using the pooled sample of children from both conditions revealed that there were no significant correlations between the ToM Scale and the CSUS (all $r(100) < 0.14$, $p > 0.17$). Nor were any significant correlations found between the Interview scores and the CSUS (all $r(100) < 0.14$, $p > 0.17$). Correlations within each condition run separately revealed the same null results.

In this first experiment, children were tested on their attribution of a Theory of Mind to robots versus humans using a direct and interactive task (ToM Scale) and an indirect task, an Interview. Children performed similarly on both tasks across conditions, attributing an equal number of mental states to the robot and the human. In line with the scalability of these items, all children, regardless of condition, performed better on the first three items than on the more difficult ones[21,25]. Overall, children's performance in the Human Condition mirrors prior work, whereas the performance of those assigned to the Robot Condition indicates less mentalizing on some items.

One unexpected difference in the performance between the conditions was found in the Diverse Desires item of the ToM Scale. Children passed this item at much higher rates in the Human Condition than the Robot Condition. One explanation for this difference is that children understood that the robot could not eat and responded with their personal snack preference. Prior work supports this interpretation with 4- and 5-year-olds stating a robotic dog lacks biological attributions, including the ability to eat[33].

The performance on the Interview mirrored that of the ToM Scale and previous research using interviews[12,34]. Regarding the three subsets of questions, children attributed Intentions and Desires as well as False Beliefs to robots at similar rates to humans. However, for the Epistemic subset, children attributed more mental states to the human than the robot. Perhaps children believe humans are more sentient than robots when asked about complex mental states (e.g., teaching, learning) as opposed to being wrong (False Belief subset). Additionally, much of the work using interviews and the AMS-Q measure specifically has been conducted with children older than our sample[12], which could explain the difference in performance on the Epistemic subset.

A limitation of the ToM Scale is that for the Robot Condition, the robots were given proper names. Using a proper name for the robot protagonists featured in the vignettes may have led children to anthropomorphize the robot. Another limitation was that the robot figurines used in the Robot Condition, despite being distinct colors, were replicas of the same robot, NAO. Study 2 addressed these limitations.

Study 2

As Theory of Mind is further developed by age 5, Study 2 aimed to investigate mental state attributions to robots in slightly older children. Although our sample in Study 2 was marginally older, we hypothesized that children would anthropomoprphize robots at a rate equivalent to the 4-year-olds tested in Study 1. In Study 2, different humanoid robots that varied in appearance were used for each ToM Scale item. This methodological change was important, as some extant literature has highlighted how morphology plays a role in children's anthropomorphism[12,19] and that the robots children interact with in everyday life vary in appearance. Additionally, this made the Robot Condition more similar to the Human Condition, where visibly distinct human figurines that varied in morphology had been used in Study 1. As the parameters between the two conditions were now more equivalent, we expected no difference between ToM Scale performance from the Human and Robot Conditions. Also, a deviation from the ToM Scale's original procedure was removed, as no proper names were assigned to the robots or the humans; instead, the experimenter labeled them as "a robot" or "a person." This change was made in order to avoid a bias toward anthropomorphism.

A second false belief item, Location False Belief, was added to the ToM Scale to better assess children's false belief reasoning. As in Study 1, we expected children to anthropomorphize the robot and did not predict a difference between the conditions on the individual items of the ToM Scale. As interviews are widely used in the literature, and to replicate the results obtained with the Interview Task, children completed the Property Projection Task. This interview assessed animacy attribution across various domains. The Property Projection Task allowed us to interview children about various items instead of just the human or the robot in the Interview task administered in Study 1. It was hypothesized that performance on the ToM Scale would match that reported in Study 1. Additionally, we predicted that children in the Property Projection Task would be unsure about the animacy of robots, as robots are unfamiliar to most children. It was expected that children, regardless of condition, would have higher scores on the ToM Scale (i.e., direct measure) than the Property Projection Task (i.e., indirect measure), as the ToM Scale provides additional context and is more interactive. As in Study 1, parents completed the CSUS.

Method

Participants

Participants were recruited from a database of past participants in the laboratory and advertisements on social media platforms. A total of 110 participants (N_{male} = 59) aged between 57 and 63 months (M = 60 months, 15 days; range = 57 months, 5 days to 63 months, 23 days) participated in the study. An additional 9 participants were tested but excluded due to sibling and/or parental interference (n = 4) and failure to complete the study (n = 3). Robot exposure in our sample (4.5%) was similar to that of Study 1. Only two children who interacted with a robot similar to those used in the study were excluded.

Power was calculated for the General Linear Model using G*Power 3.1. The required power was estimated using the linear multiple regression: fixed model, R^2 deviation from zero analysis. Effect size was estimated at medium 0.15, alpha error probability was entered at 0.05, power at 0.80, and the number of predictors entered were 5 (Condition, Age, SES, CSUS, and Robot Exposure). For this analysis, a total sample size of 92 is required. For the ANOVA, a ANOVA repeated measures, within-between interaction power analysis was run in G*Power 3.1. Effect size was estimated at medium 0.25, alpha error probability at the standard 0.05, power at 0.80, with 5 groups (Human, Humanoid Robot, Non-humanoid Robot, Toy Car, Rodent) and 5 measurements (one score for each item listed above). The correlation among measures was set at 0.5, and the nonsphericity correction was set at 1. For this analysis, a total sample size of 35 is required. Therefore, our sample size is sufficient.

Based on their language proficiency, 23 participants were tested in French, and 87 participants were tested in English. All participants resided in either Canada or the United States. Participants identified as Caucasian (34.7%), Asian (17.3%), mixed ethnicity (18.7%), other, or did not specify (29.3%). The participants in the current study were predominantly from families with high socioeconomic status, with 40% of families reporting an annual income exceeding $100,000 and 32.7% making between $50,000 and $100,000. The remainder of the sample reported an income below $50,000 or opted not to report their family income.

Measures

The CSUS was administered as described in Study 1. The ToM Scale was administered with one additional false belief item. In lieu of the Interview Task, a Property Projection Task was used. The Robot Exposure Questionnaire was edited as described below. New measures are described below in detail.

Robot Exposure Questionnaire

The Robot Exposure Questionnaire was modified slightly from Study 1 to improve the range of robot exposure that could be captured. For this purpose, most questions were changed from a binary yes/no response to a Likert Scale. Questions included (1) whether the family possessed a robot at home, (2) children's frequency of interacting with robots both at home and (3) outside of the home (e.g., school, friend/family member's home), and (4) how often children play video games or watch TV/movies that feature robot characters. Question 1 was a yes/no/unsure forced response. Questions 2 through 4 were on a 5-point Likert scale from "Never" to "Very Often."

Property Projection Task

As interviews are widely used in the literature, the Property Projection Task was used in Study 2. Additionally, this task allowed for the assessment of multiple domains and items as outlined below. The Property Projection Task included 8 questions, 2 per domain[17]. This task provides a comprehensive understanding of mental states across a variety of domains: Artifact (e.g., "Did a person make this robot/rat/car/person?"), sensory (e.g., "If I tickled this robot/rat/car/person, would this robot/rat/car/person feel it?"), biological (e.g., "Does this robot/rat/car/person eat?"), and psychological (e.g., "Can this robot/rat/car/person feel happy?")[17]. Participants responded to the questions with "yes" or "no." In the present study, children were shown images of a humanoid and a non-humanoid robot, a rodent, a toy car, and a human on the screen (see Fig. 4). Note that the image of the human was gender-matched to the child in Study 2. For a full list of questions, please see Appendix C.

Theory of Mind Scale

The Wellman and Liu Task was administered as in Study 1, with the addition of a 6th item, a change of Location False-Belief task. For this additional item, a classic Sally and Anne paradigm was employed. For this item, a ball was placed in a blue box, the human or robot figurine left the scene, and the experimenter moved the ball from the blue box (original location) to the pink box (new location). To be awarded a point for this item, children had to respond that the robot or human would look for the item in its original location, explain that the ball was

Fig. 4. The images used for the Property Projection Task.

actually in the new location, and answer "no" when asked whether the robot/human had moved the ball. This new item was added to better assess children's ToM skills, as the Sally-Anne paradigm is widely used in other research. Furthermore, the robot figurines used varied in appearance (but were all humanoid) to more closely match the varied human figurines (see Fig. 5). The human figurines were identical to those in Study 1, and images identical to those shown in Table 1 were used for the scale.

Materials

The materials used in Study 2 were identical to those used in Study 1 except that for the ToM Scale, five 3D-printed humanoid robot figurines, white in color with black accents to highlight key features, were used. The humanoid robots were chosen from the anthropomorphic robot (ABOT) database[35], with scores varying slightly in human likeness. Specifically, humanoid robots were chosen with human-likeness scores varying from 42.17 (Pepper) to 51.26 (Kirobo). In the Location False Belief item, two boxes (2x2 inches) that differed in color (i.e., blue and pink) and a small ball were used. Digital images of a humanoid and non-humanoid robot, person, toy car, and rodent were presented on a screen for the Property Projection Task.

Procedure

The procedure of Study 2 mirrored that of Study 1. The Robot Exposure Questionnaire and the other forms described in Study 1 were sent to the parent electronically and completed before the study session. The Welcome PowerPoint, the parent's verbal consent, and the child's verbal assent were identical to Study 1. Children were randomly assigned to one of two conditions: Human or Humanoid Robot for the ToM Scale. In the Human Condition, human figurines were used for the ToM Scale, whereas humanoid robot figurines were used in the Robot Condition. Before responding to subsequent questions within each item of the scale, the participants were shown the figurine and asked, "What is this?". If the child incorrectly labeled the figurine as anything other than a human or robot, the experimenter provided the correct label. This ensured that all participants had the same understanding of each figurine (i.e., whether it was a human or robot protagonist). All participants completed the Property Projection Task and the ToM Scale in a counterbalanced order. The order in which the figurines were used in the scale was also counterbalanced. As an additional item, the Location False Belief item was added to the ToM Scale, and the order of the false belief items (location and contents) was counterbalanced. In the Property Projection Task, the order in which the items (human, humanoid robot, non-humanoid robot, toy car, rodent) were presented was also counterbalanced. The debrief procedure at the conclusion of the study was identical to Study 1.

Scoring

Children received a passing or failing score for each ToM Scale item. Therefore, scores on this task ranged from 0 to 6. For the Property Projection Task, a positive answer yielded one point. The two artifact questions ("Can this break?" and "Did a person make this?") were reverse-coded so that higher scores would indicate more anthropomorphism across all questions. There were 8 questions per item, and total scores for all items ranged from 0 to 40. Each child received a mean total score for the CSUS (out of 4). For the Robot Exposure Questionaire, any data cells where parents had answered "unsure" were left blank. Otherwise, children received scores based on their robot exposure, as reported by the parent. For question one, 'yes' was coded as 4 (to match the Likert Scale) and 'no' as 0. For the rest, scores ranged from 0 (Never) to 4 (Very Often). The score for each of the 4 questions was averaged to create an overall Robot Exposure Score per child, with higher scores indicating more familiarity with and experiences with robots. To categorize Socio-Economic Status as a variable, SES was split into 3 categories, with families making less than $50,000 classified as low SES, $50,0000 to $100,000 classified as mid, and over $100,000 classified as high.

Results and Discussion

Data Analysis

The data analysis plans matched those of Study 1, with the Property Projection Task replacing the Interview Task. Unless otherwise specified, statistical analyses were performed using JASP 0.18.3[30].

Theory of Mind Scale

Children in the Human Condition ($M = 3.44$, $SD = 1.41$; $t(54) = 2.29$, $p = 0.026$, $d = 0.31$) and the Robot Condition ($M = 3.73$, $SD = 1.39$; $t(54) = 3.87$, $p < 0.001$, $d = 0.52$) performed as expected (above chance) on the ToM Scale. A General Linear Model, run in R Version 4.3.3[31] and using the *glm* function, compared ToM Scale performance across conditions (Robot and Human). Age (in months), Socio-Economic Status (Low, Mid, or High), CSUS score, and overall Child Robot Exposure (as measured by the exposure questionnaire) were entered as factors.

Fig. 5. The robot figurines used in the Theory of Mind Scale for Study 2.

ToM scale item	Human condition	Robot condition	Difference between the human and robot conditions
Diverse Desires	87.3***	89.1***	$X^2(1, N=110)=0.09, p=0.77$
Diverse Beliefs	78.2***	81.8***	$X^2(1, N=110)=0.23, p=0.63$
Knowledge Access	65.5*	72.7**	$X^2(1, N=110)=0.68, p=0.41$
Contents False Belief	43.6	47.3	$X^2(1, N=110)=0.15, p=0.70$
Change of Location	43.6	54.5	$X^2(1, N=110)=1.31, p=0.25$
Hidden Emotion	25.5***	27.3**	$X^2(1, N=110)=0.05, p=0.83$

Table 4. Mean percentage of children who passed each condition on the Theory of Mind Scale. * Represents a *p*-value of less than 0.05, ** below 0.01, *** below 0.001 in comparison to chance.

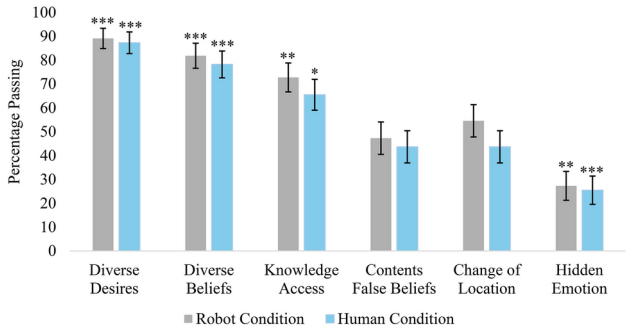

Fig. 6. Percentage of children who passed each item of the Theory of Mind Scale. Error bars represent standard error. * Represents a *p*-value of less than 0.05, ** below 0.01, *** below 0.001 compared to chance.

Item	Mean	SD	T-test
Human	7.19***	1.13	$t(109)=29.64, p<0.001, d=2.83$***
Rodent	5.92***	1.50	$t(109)=13.44, p<0.001, d=1.28$***
Humanoid Robot	3.09***	1.70	$t(109)=-5.63, p<0.001, d=-0.54$***
Non-humanoid Robot	3.05***	1.74	$t(109)=-5.76, p<0.001, d=-0.55$***
Car	1.87***	1.61	$t(109)=-13.87, p<0.001, d=-1.32$***

Table 5. Performance per Item on the Property Projection Task . *** Represents a *p*-value of less than 0.001 in comparison to chance.

Condition was found not to be significant (X^2 (1) = -0.28, $p=0.31$), indicating no difference in performance between the Human and Robot Conditions. Age, however, had a significant effect (X^2 (1) = 0.15, $p=0.02$), with older children performing better. All other factors did not reach significance. The overall R^2 for this model was 0.1. Next, each item of the ToM Scale was analyzed separately to determine potential condition effects within the task.

A series of binomial tests revealed that most children in both the Human and Robot Conditions succeeded on the Diverse Desires, Diverse Beliefs, and Knowledge Access items but failed on the Hidden Emotion, Contents False Belief, and Change of Location items. Children in the Human Condition and children in the Robot Condition performed similarly on all items in the scale; see Table 4 and Fig. 6.

Property Projection Task

When looking at all the questions combined, children performed well on the Property Projection Task, attributing animacy to the human and the rodent and not attributing animacy to the robots and the car, see Table 5. When looking at only the sensory and psychological questions, children still performed well on most items. They attributed psychological and sensory characteristics to the human and the rodent and did not attribute them to the car. However, they were confused about the psychological and sensory capabilities of robots, performing at chance for both the humanoid and non-humanoid robots, see Table 6.

A repeated-measures analysis of variance test (ANOVA) compared the 5 items total scores to each other. Mauchly's test of sphericity indicated that sphericity was violated, therefore a Greenhouse–Geisser correction was applied to the model. The ANOVA yielded a significant main effect ($F(2.98, 324.85)=275.72$, $p<0.001$, $n^2_p=0.72$), indicating different performances across items. Post-hoc tests revealed that all items significantly differ in performance from one another except the total score of humanoid robots versus non-humanoid robots

Item	Mean	SD	T-test
Human	3.79***	0.56	$t(109) = 33.54$, $p < 0.001$, $d = 3.20$***
Rodent	2.82***	1.18	$t(109) = 7.26$, $p < 0.001$, $d = 0.69$***
Humanoid Robot	2.08	1.25	$t(109) = 0.69$, $p = 0.49$, $d = 0.07$
Non-humanoid Robot	1.94	1.38	$t(109) = -0.48$, $p = 0.63$, $d = -0.05$
Car	0.81***	1.21	$t(109) = -10.34$, $p < 0.001$, $d = -0.99$***

Table 6. Performance per Item on the Property Projection Task: Psychological and Sensory Questions. *** Represents a p-value of less than 0.001 in comparison to chance.

($t(109) = 0.37$, $p_{holm} = 0.72$). An ANOVA looking at only the sensory and psychological scores combined revealed the same pattern of results.

Children's Social Understanding Scale (CSUS)
All parents filled out the CSUS, resulting in no missing data[28]. Participants' total CSUS score was calculated ($M = 3.14$, $SD = 0.36$). Overall, parents rated their children as having levels of Theory of Mind in line with those reported in prior research at a similar age[28,29].

Exploratory Inter-tasks Comparison
Unlike in Study 1, correlations were run only on a full sample and not split by condition. This was done because the children in the Robot and Human Conditions were both asked about all 5 items for the Property Projection Task, resulting in no differences for this task. The ToM Scale was negatively correlated with the most and least animate items of the Property Projection Task total scores, the car ($r(108) = -0.28$, $p = 0.003$) and the human ($r(108) = 0.20$, $p = 0.035$). Therefore, better performance on the ToM Scale correlated with worse performance on the Human and Car items of the Property Projection Task total scores. When looking at only the sensory and psychological combined scores of the Property Projection Task, the human ($r(108) = 0.25$, $p = 0.009$) and rodent ($r(108) = 0.20$, $p = 0.03$) items were positively correlated with ToM Scale Performance. Therefore, higher attribution of animacy to living beings correlated with better performance on the ToM Scale. The Property Projection Task was not significantly correlated with the CSUS using either the total scores or the sensory/psychological scores. However, the ToM Scale was trending towards a positive correlation with the CSUS ($r(108) = 0.18$, $p = 0.06$). When broken down by ToM Scale condition, the Robot Condition is not significantly correlated with the CSUS ($r(53) = 0.08$, $p = 0.58$) but the Human Condition ($r(53) = 0.36$, $p = 0.007$) is significantly correlated with the CSUS. Therefore, better performance on the CSUS predicts better performance on the ToM Scale when administered with human figurines.

The findings generally mirrored those of Study 1. Specifically, children attributed mental states to humans and robots at similar rates when measured with the ToM Scale on all the items. Regardless of condition, children's performance on each item of the scale matched previous studies[21,25]. Unlike in Study 1, there were no condition differences for any of the items. Perhaps using humanoid robot figurines that varied in appearance better matched the human figurines, leading to more similar performance across the conditions. The increased age of our sample may have also helped them understand the demands of the task. To further investigate the role that morphology could play on children's ToM attribution, future work should administer the ToM Scale with non-humanoid robot figurines that vary in their morphological appearance.

Regarding the Property Projection Task, children correctly attributed animacy to the human and the rodent. They were also correct in judging the toy car as inanimate. Looking at the total score for the Property Projection Task confirms that 5-year-olds knew that both the humanoid and non-humanoid robots were inanimate. However, this finding was likely driven by the biological questions, as children were confused about the psychological and sensory properties of both robots. Children's knowledge of the biological properties of robots at age 5 mirrors prior work with interviews and experimental tasks[19].

Overall, results suggest that children's attribution of mental states is not limited to humans but extends to robots as well when investigated with experimental tasks that provide context for reasoning. Without such context, as in interviews, children are confused about the sensory and psychological properties of social robots.

General Discussion

The present work aimed to examine young children's attribution of mental states to humanoid robots in a novel way. In contrast to prior work that used interviews to assess children's mental state attribution, we tested children with an experimental task[19]. To our knowledge, this is the first study to use the entire ToM Scale with non-human protagonists. Additionally, the present work contrasted direct (experimental) and indirect (interviews) measures to assess children's Theory of Mind skills, allowing for a direct comparison of anthropomorphizing as a function of the richness of context.

Taken together, the results of the present studies show that 4- and 5-year-old children attribute mental states to humanoid robots and humans similarly when measured with the ToM Scale. In contrast, when tested with the Interview Task (Study 1), children attributed more epistemic properties to humans. One explanation for this difference between the Human and Robot conditions could be the rich context (i.e. vignettes of familiar daily situations) provided in the ToM Scale. As such, the robots are treated as depictions of social agents more easily than in the context of an interview[36]. The difference in the attribution of epistemic properties between the Human and Robot conditions in the interview format could be due to mental state terms (e.g., think, know)

being associated with humans in conversations that are more likely to impact verbal questioning. In Study 2, a similar pattern of results was observed with a different interview and methodological improvements in the administration of the scale. Notably, we reported a replication of the ToM Scale with slightly older children, with 5-year-olds attributing mental states equally to humans and humanoid robots. Again, the Property Projection Task, also an interview, reflected less mentalizing to robots than to humans. This suggests that when no human-like context is provided in an Interview task, it is more challenging for young children to anthropomorphize robots.

Another possibility for the contrasting results when using experimental (i.e., the ToM Scale) and non-experimental tasks (i.e., interviews) could be that the realistic vignettes used for each item made children artificially anthropomorphize the protagonists by automatically assigning them certain animacy characteristics (e.g., preferring one snack in the Diverse Desires item). Future research should investigate mentalizing with the ToM Scale with a range of inanimate protagonists (e.g., non-humanoid robots, sticks) to confirm this hypothesis. Furthermore, the ToM Scale should be administered to older children with higher levels of ToM.

Regarding the link between Theory of Mind skills (CSUS) and mentalizing in experimental tasks, no significant correlations were observed with the interview measures of mentalizing in both studies, replicating previous work with robots[11]. In contrast, parental reports of the child's ToM skills with the CSUS were positively linked to performance on the Scale in the Human Condition in Study 2. This might suggest that the more concrete assessment of ToM with daily situations depicted in vignettes, as used in the ToM scale, might be a more ecologically valid measure of ToM skills in children. Why such a link was not observed when robots replaced the human figurines might be explained by the fact that human figurines facilitated reasoning about mental states. However, it is important to point out that research findings are mixed regarding whether the CSUS scores are concurrently and positively associated with behavioral ToM tasks[28,37,38].

The present work suggests a number of avenues for future research. In Study 1, figurines representing the same robot, NAO, were used in the ToM Scale. Study 2 improved on this design by using humanoid robots that differed in morphology for each scale item. Future work could use the ToM Scale to investigate children's mental state attribution to non-humanoid robots. If no differences are observed across non-humanoid and humanoid robots, one would conclude that morphology may not be a driving factor in children's mentalizing of robots and that the "human" context provided by the ToM Scale triggers anthropomorphism. Some studies have shown that the robot's morphology can impact children's mental state attribution[11,12]. Although morphology did not impact mentalizing in the present work, the role of morphology in children's attribution of mental states to robots plays an important role in some contexts and warrants further investigation.

Additionally, future research could examine how children's prior exposure to robots (e.g., live interactions, with books or television) could impact their mental state attribution. The current studies tested children for whom robots were unfamiliar, but future work could test a group of children with regular exposure to robots. Additionally, asking children to justify their responses to interview questions could provide valuable insights into young children's perceptions of robots, potentially informing the design and use of robots in educational and other settings. Finally, future work should identify the cognitive mechanisms involved in these observed developmental changes. As the development of ToM is complex and multifaceted, understanding which mechanisms play a role in ToM development can help us explain why children anthropomorphize and how this could impact their emotional understanding, empathy towards others, and their views on morals/values concerning non-human social agents.

Data availability

Data is available by request. Correspondence and requests for data and materials should be addressed to E.J.G.

Received: 12 August 2024; Accepted: 24 March 2025
Published online: 10 May 2025

References

1. Airenti, G. The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Front. Psychol.* **9**, 2136 (2018).
2. Hatano, G. et al. The development of biological knowledge: A multi-national study. *Cogn. Dev.* **8**(1), 47–62. https://doi.org/10.1016/0885-2014(93)90004-O (1993).
3. Ochiai, M. The role of knowledge in the development of the life concept. *Hum. Dev.* **32**(2), 72–78. https://doi.org/10.1159/000276365 (1989).
4. Opfer, J. E. & Gelman, S. A. Development of the animate-inanimate distinction. In *The Wiley-Blackwell handbook of childhood cognitive development* (ed. Goswami, U.) 213–238 (Wiley Blackwell, 2011).
5. Piaget, J. *The Child's Conception of the World* (Kegan Paul, Trench & Trubner, 1929).
6. Poulin-Dubois, D. & Héroux, G. Movement and children's attributions of life properties. *Int. J. Behav. Dev.* **17**(2), 329–347. https://doi.org/10.1177/016502549401700206 (1994).
7. Rakison, D. H. & Poulin-Dubois, D. Developmental origin of the animate–inanimate distinction. *Psychol. Bull.* **127**(2), 209–228. https://doi.org/10.1037/0033-2909.127.2.209 (2001).
8. Richards, D. D. & Siegler, R. S. Children's understandings of the attributes of life. *J. Exp. Child Psychol.* **42**(1), 1–22. https://doi.org/10.1016/0022-0965(86)90013-5 (1986).
9. Di Dio, C. et al. Shall I trust you? from child-robot interaction to trusting relationships. *Front. Psychol.* https://doi.org/10.3389/fpsyg.2020.00469 (2020).
10. Dunham, P., Dunham, F., Tran, S. & Akhtar, N. The nonreciprocating robot: Effects on verbal discourse, social play, and social referencing at two years of age. *Child Dev.* **62**(6), 1489–1502. https://doi.org/10.2307/1130821 (1991).
11. Goldman, E. J., Baumann, A. E. & Poulin-Dubois, D. Preschoolers' anthropomorphizing of robots: Do human-like properties matter?. *Front. Psychol.* **13**, 8708. https://doi.org/10.3389/fpsyg.2022.1102370 (2023).
12. Manzi, F. et al. A robot is not worth another: Exploring children's mental state attribution to different humanoid robots. *Front. Psychol.* https://doi.org/10.3389/fpsyg.2020.02011 (2020).

13. Okanda, M., Taniguchi, K., Wang, Y. & Itakura, S. Preschoolers' and adults' animism tendencies toward a humanoid robot. *Comput. Hum. Behav.* **118**, 106688. https://doi.org/10.1016/j.chb.2021.106688 (2021).
14. Poulin-Dubois, D., Lepage, A. & Ferland, D. Infants' concept of animacy. *Cogn. Dev.* **11**(1), 19–36. https://doi.org/10.1016/S0885-2014(96)90026-X (1996).
15. Saylor, M. M., Somanader, M., Levin, D. T. & Kawamura, K. How do young children deal with hybrids of living and non-living things: The case of humanoid robots. *Br. J. Dev. Psychol.* **28**(4), 835–851. https://doi.org/10.1348/026151009X481049 (2010).
16. Baumann, A. E., Goldman, E. J., Meltzer, A. & Poulin-Dubois, D. People do not always know best: preschoolers' trust in social robots. *J. Cogn. Dev.* https://doi.org/10.1080/15248372.2023.2178435 (2023).
17. Jipson, J. L., Gülgöz, S. & Gelman, S. A. Parent-child conversations regarding the ontological status of a robotic dog. *Cogn. Dev.* **39**, 21–35. https://doi.org/10.1016/j.cogdev.2016.03.001 (2016).
18. Manzi, F., Massaro, D., Kanda, T., Kanako, T., Itakura, S., & Marchetti, A. Teoria della Mente, bambini e robot: l'attribuzione di stati mentali [Theory of Mind, children, and robots: The attribution of mental states]. In *XXX Congresso AIP Sezione di Psicologia dello Sviluppo e dell'Educazione, Messina, Italy.* (2017).
19. Goldman, E. J. & Poulin-Dubois, D. Children's anthropomorphism of inanimate agents. *Wiley Interdisc. Rev. Cogn. Sci.* **15**, e1676 (2024).
20. van Straten, C. L., Peter, J. & Kühne, R. Child-robot relationship formation: A narrative review of empirical research. *Int. J. Soc. Robot.* **12**, 325–344. https://doi.org/10.1007/s12369-019-00569-0 (2020).
21. Hiller, R. M., Weber, N. & Young, R. L. The validity and scalability of the Theory of Mind Scale with toddlers and preschoolers. *Psychol. Assess.* **26**(4), 1388–1393. https://doi.org/10.1037/a0038320 (2014).
22. Kuntoro, I. A., Peterson, C. C. & Slaughter, V. Culture, parenting, and children's theory of mind development in Indonesia. *J. Cross Cult. Psychol.* **48**(9), 1389–1409. https://doi.org/10.1177/0022022117725404 (2017).
23. Peterson, C. C., Wellman, H. M. & Slaughter, V. The mind behind the message: advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Dev.* **83**(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x (2012).
24. Poulin-Dubois, D., Goldman, E. J., Meltzer, A. & Psaradellis, E. Discontinuity from implicit to explicit theory of mind from infancy to preschool age. *Cogn. Dev.* https://doi.org/10.1016/j.cogdev.2022.101273 (2023).
25. Wellman, H. M., Fang, F., Liu, D., Zhu, L. & Liu, G. Scaling of theory-of-mind understandings in Chinese children. *Psychol. Sci.* **17**(12), 1075–1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x (2006).
26. Zhang, Y. et al. Theory of robot mind: False belief attribution to social robots in children with and without autism. *Front. Psychol.* https://doi.org/10.3389/fpsyg.2019.01732 (2019).
27. Wellman, H. M. & Liu, D. Scaling of theory-of-mind tasks. *Child Dev.* **75**(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x (2004).
28. Tahiroglu, D. et al. The children's social understanding scale: construction and validation of a parent-report measure for assessing individual differences in children's theories of mind. *Dev. Psychol.* **50**(11), 2485–2497. https://doi.org/10.1037/a0037914 (2014).
29. Brosseau-Liard, P. & Poulin-Dubois, D. Fiabilité et validité de l'Échelle de compréhension sociale des enfants. *Psychol. Fr.* **64**(4), 331–341 (2019).
30. JASP Team. JASP (Version 0.18. 3)[Computer software]. (2024).
31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (2022).
32. The jamovi project. jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org. (2022).
33. Jipson, J. L. & Gelman, S. A. Robots and rodents: Children's inferences about living and nonliving kinds. *Child Dev.* **78**(6), 1675–1688 (2007).
34. Miraglia, L. et al. Development and validation of the attribution of mental states questionnaire (AMS-Q): A reference tool for assessing anthropomorphism. *Front. Psychol.* **14**, 999921. https://doi.org/10.3389/fpsyg.2023.999921 (2023).
35. Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (105–113). (2018).
36. Clark, H. H. & Fischer, K. Social robots as depictions of social agents. *Behav. Brain Sci.* **46**, e21. https://doi.org/10.1017/S0140525X22000668 (2022).
37. Akbulut, M., Etel, E., Tahiroğlu, D. & Selçuk, A. B. Children's social understanding scale-short form: Adaptation to Turkish sample. *Early Educ. Dev.* **34**(1), 329–347 (2023).
38. Poulin-Dubois, D., Goldman, E. J., Meltzer, A. & Psaradellis, E. Discontinuity from implicit to explicit theory of mind from infancy to preschool age. *Cogn. Dev.* **65**, 1–15. https://doi.org/10.1016/j.cogdev.2022.101273 (2023).

Acknowledgements

We thank Mihaela Zlatanovska for her help with scheduling and recruitment. We extend our thanks to the families who participated in the study.

Author contributions

E.J.G, A-E.B., and D.P-D. conceived the experiments. L.P. and J.B. conducted the experiments. E.J.G and A-E.B. analyzed the results. E.J.G, A-E.B., L.P., J.B., and D.P-D. contributed to the writing and revision of the manuscript.

Funding

This work was supported by an Insight Grant from the Social Sciences and Humanities Research Council of Canada (#435-2022-0805), which was awarded to D.P-D.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96229-7.

Correspondence and requests for materials should be addressed to E.J.G.