Cognition 254 (2025) 105983

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Developmental changes in the perceived moral standing of robots

Madeline G. Reinecke [a,*], Matti Wilks [b], Paul Bloom [c,d]

[a] University of Oxford, United Kingdom
[b] University of Edinburgh, United Kingdom
[c] Yale University, USA
[d] University of Toronto, Canada

ARTICLE INFO

Keywords:
Moral cognition
Developmental psychology
Artificial intelligence
Human-AI interaction

ABSTRACT

Emerging evidence suggests that children may think of robots—and artificial intelligence, more generally—as having moral standing. In this paper, we trace the developmental trajectory of this belief. Over three developmental studies (combined $N = 415$) and one adult study ($N = 156$), we compared participants' judgments (Experiments 1–3) and donation choices (Experiment 4) towards a human boy, a humanoid robot, and control targets. We observed that, on the whole, children endorsed robots as having moral standing and mental life. With age, however, they tended to deny experiential mental life to robots, which aligned with diminished ascription of moral standing. Older children's judgments more closely mirrored those of adult participants, who overwhelmingly denied these attributes to robots. This sheds new light on children's moral cognitive development and their relationship to emerging technologies.

In 2016, Hanson Robotics unveiled a humanoid robot named Sophia—designed to be a "unique combination of science, engineering, and artistry" (Hanson Robotics, 2023). In the years following her release, Sophia accumulated a long list of speaking engagements and accolades (ranging from television appearances to sitting on panels regarding the "social good" of artificial intelligence; NBC4 Washington, 2023). One of these honors, however, is especially striking: In October 2017, Sophia was awarded citizenship in Saudi Arabia. This made her the first instance of artificial intelligence (AI) to have legal standing (and, debatably, moral standing) equivalent to that of a human (Parviainen & Coeckelbergh, 2021). To be sure, Sophia's citizen status is purely a legal designation—but these kinds of advances raise a host of questions within the moral domain. Would turning Sophia off be morally equivalent to killing a human? Does Sophia deserve the same kind of moral treatment afforded to persons?

This paper sets aside normative matters, such as whether instances of artificial intelligence *actually* have moral standing (or whether they will in the future). We instead take up the descriptive question: What do people think about the moral standing of artificial intelligence, and how do these beliefs form?

1. Developmental focus

We approach the study of AI moral standing from a developmental perspective, comparing the judgments of children and adults. This serves two purposes.

First, we gain insight into not only what people's moral standing beliefs are, but how these beliefs might form. Just as children (Jordan et al., 2014) and adults alike appear predisposed to favor those similar to them (Chae et al., 2022), they may similarly be predisposed to privilege the moral standing of humans over non-humans, like AI. Alternatively, children and adults could disagree about the moral standing of artificial minds. This would instead suggest that our moral standing beliefs are malleable and learned over the course of development.

Second, we take children's beliefs about the moral standing of artificial minds as worthy of investigation in and of themselves. Children today will grow up with a closer relationship to technology than ever before. Understanding their perspectives on artificial intelligence may forecast the nature of human-computer social and moral interaction.

2. Considerations relevant to judgments of moral standing

One consideration relevant to people's judgments of moral standing is category membership. As sophisticated as humanoid robots may be,

* Corresponding author.
E-mail addresses: madeline.reinecke@psych.ox.ac.uk (M.G. Reinecke), mwilks@ed.ac.uk (M. Wilks), paul.bloom@utoronto.ca (P. Bloom).

https://doi.org/10.1016/j.cognition.2024.105983
Received 11 June 2024; Received in revised form 5 October 2024; Accepted 9 October 2024
Available online 9 November 2024

they might not qualify for the same moral status as humans, merely because they are not humans. In research pitting human interests against the interests of non-human animals (e.g., chimpanzees), adults tended to prioritize human moral standing even when other factors (e.g., intelligence) were held constant (Caviola et al., 2022). Though the majority of this literature focuses on distinguishing humans versus non-human animals, this consideration of "humanness" may similarly apply to considering the moral worth of robots and artificial minds (Nijssen et al., 2019). People with speciesist attitudes may place AI—along with many forms of non-human animals—in the outer ranks of the moral circle.

Such speciesist moral beliefs may emerge over time (e.g., Neldner et al., 2018). Between the ages of 5 and 9, children prioritize saving humans (at the expense of non-human animals) less often than do adults (Paruzel-Czachura et al., 2024; Wilks et al., 2021), and they care more about the moral concerns of (at least some) robots (Sommer et al., 2019). Indeed, children around 3-years-old even engage in spontaneously helping behavior towards humanoid robots (Martin et al., 2020)—mirroring similar findings regarding early-emerging prosocial behavior towards humans (Warneken & Tomasello, 2009).

Another consideration relevant to the determination of moral standing is how people perceive the mental life of different entities (e.g., Gray et al., 2007; Ladak, 2024). If artificial intelligence is seen as capable of suffering, this could bear important moral implications (Shevlin, 2021). Indeed, ex–Google engineer Blake Lemoine became convinced that the large language model "laMDA" was sentient and therefore being deprived of its rights (Tiku, 2022).

Do people's beliefs about the moral standing of AI hinge on how they judge a given system's mental capacities? In a landmark paper, Gray et al. (2007) observed that people tie morality with mind: Participants rated targets with greater "agency," or capacity for higher-order mental life, as deserving blameworthiness for their actions, whereas targets with greater "experience," or capacity for phenomenal states, deserved greater protection from harm. These relationships were observed across more than a dozen targets, including humans across the lifespan, chimpanzees, dogs, frogs, God, and—importantly for our purposes—robots. Adults evaluated the robot target as having some degree of agency but little experience, suggesting that robots may be held accountable for misdeeds (but have limited moral standing themselves).

Young children, however, appear more willing than older children and adults to ascribe mental abilities to artificial minds (Brink et al., 2019; Kahn et al., 2012; Manzi et al., 2020; Weisman et al., 2017). This may correspond with beliefs about moral standing. Children seem more likely than adults to grant robots with "socio-emotional capacities," like being able to feel love (Weisman et al., 2017), and children who grant robots with greater mental capacities are less likely to see them as "creepy" (at least before age 9; Brink et al., 2019). This tendency to assign robots with mental and moral characteristics seems to diminish over development. After witnessing an experimenter transgressing against a robot during a lab visit, for example, 15-year-olds proved less likely to see the robot as a "mental and moral other" than 9-year-olds and 12-year-olds (Kahn et al., 2012). As it stands, though, this existing literature cannot address how specific aspects of mind perception, like agency and experience, align with children's beliefs about robot moral standing (or whether these beliefs shift in tandem over development).

In this paper, we explore the relationship between mind perception and moral status—expanding on this existing developmental literature (e.g., Flanagan et al., 2023; Kahn et al., 2012; Sommer et al., 2019; Weisman et al., 2017). Specifically, the present research stands to provide deeper understanding of how specific aspects of mind perception (e.g., agency, experience) relate to moral standing judgments in early to middle childhood. In four experiments, we examine children and adults' beliefs about the moral standing of a robot, human, toy bear (control, Experiments 1–4), and a rock (additional control, Experiment 4), alongside their beliefs about the mental lives of these entities. From this, we hope to gain new understanding of how moral standing judgments form within the domain of artificial intelligence, as well as whether these judgments draw on inferences related to mind perception. We see this as informative for understanding (and potentially even forecasting) human-AI social interaction.

For access to our preregistrations (when applicable), materials, data, and R scripts across all studies, please see the Open Science Framework. All studies received ethical approval from the Yale University Psychology Department's Institutional Review Board. For developmental studies, participants' parents provided consent (and children provided assent) before the start of the research. For adult research, participants consented before participating.

3. Experiment 1

For this exploratory study, we investigated how children's evaluations of mental life and moral standing might differ across different kinds of entities. To examine these effects, we provided children with stories about a human boy transgressing against another human, a robot, or a toy bear. The toy bear served as a control, allowing us to identify whether children mentalize (and moralize) robots differently from artifacts commonly anthropomorphized during early development (e.g., toys). Participants evaluated the extent to which their target was harmed by these transgressions, whether the agent intended to perform the transgression, and whether the agent deserved punishment. In another phase, participants evaluated the mental capacities of the target victim.

3.1. Method

3.1.1. Participants

We collected data in-person from 123 children between the ages of 4 and 13. Again, we treated this as an exploratory investigation—meaning that we did not preregister hypotheses or analyses, set an *a priori* stopping rule, or evenly sample based on age. We removed data from seven participants prior to analysis for either (1) failing a comprehension check ($n = 6$), or (2) not having age-related demographic information attached to their data ($n = 1$). This left 116 children in our final sample ($M_{age} = 7.73$, $SD_{age} = 2.09$; 59 identified as male, 57 identified as female).

3.1.2. Materials and procedure

We randomly assigned participants to either the human boy, robot, or toy bear condition. We described each event identically across these conditions, except for target-based information (e.g., "This is Drew. He's a [boy/robot/toy bear]").

In the transgression phase of the experiment, each participant evaluated one physical transgression and one non-physical transgression (in counterbalanced order) directed towards their target (see Fig. 1). After hearing about the transgression, we asked participants, "Do you think being [pushed onto the ground / called a mean name] hurt Drew?" (Yes/No). If participants said yes, we then asked them "how much" they thought the target was hurt (a teeny bit, a little bit, or a lot). Using this same metric, we also asked about the agent's intentionality (i.e., "Do you think Matt meant [to push Drew onto the ground]?") and whether the agent deserved punishment (i.e., "Do you think Matt should get in trouble for [pushing Drew onto the ground]?").

In the "mental life" phase, each participant evaluated their target on four higher-order mental capacities (i.e., self-control, memory, communication, planning) and four experiential mental capacities (i.e., anger, fear, hunger, happiness), drawn from Gray et al. (2007). Mirroring the transgression phase of the experiment, we asked each of these questions in a yes-no format (e.g., "Do you think Drew can feel angry?"), followed by a three-point scale (if participants responded affirmatively) to gauge the strength of their responses. If a child affirmed that a robot could feel happy, for example, the experimenter would then ask, "How much do you think Drew can feel happy—a teeny bit, a little bit, or a lot?". We then ended the experiment with a comprehension check and

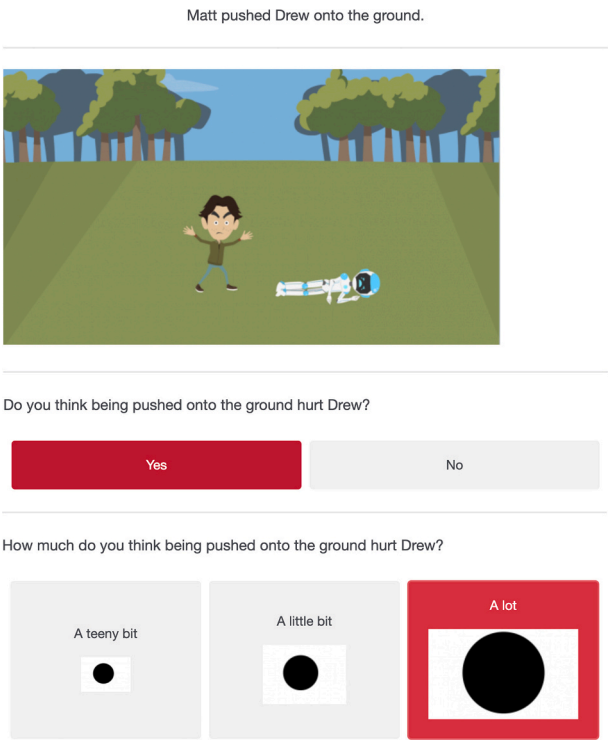

Fig. 1. Example stimuli used in Experiments 1 and 2. © 2024 GoAnimate, Inc. Images are copyrighted by and used by permission of VYOND™. VYOND is a trademark of GoAnimate, Inc., registered in Australia, Brazil, the European Union, Norway, the Philippines, Singapore, Switzerland and the United Kingdom.

provided the child with a prize. We presented all stimuli to participants using an iPad with the iOS Qualtrics application.

3.2. Results & discussion

In the following subsections, we investigate children's evaluations of (1) robot mental life, (2) robot suffering, (3) punishment deserved by a human agent after transgressing against a robot, and (4) the intentionality of a human agent who transgressed against a robot. Given that we did not sample children evenly by age for this experiment (and testing for age effects would violate model assumptions), we collapsed across all participants for these analyses.

3.3. Data preparation

To determine the degree of mental life, harm vulnerability, punishment, and intentionality that children endorsed on each trial, we created a continuous scale by re-coding "No" responses as 0, "a teeny bit" as 1, "a little bit" as 2, and "a lot" as 3 (as has become standard within related areas of developmental psychology; e.g., Flanagan et al., 2023). To see our analysis of the non-re-coded data (i.e., the binary judgments across all metrics)—which echoes the analyses presented in the primary text—see the "Supplemental Online Materials" file on the Open Science Framework.[1]

[1] There are a few discrepancies to report between the supplementary analysis (i.e., regarding binary data) and the primary analysis (i.e., regarding continuous data). Specifically, children's binary evaluations of harm vulnerability did not differ across targets, nor did children distinguish between the punishment deserved by transgressors across targets. For more detail, please see the SOM.

3.4. Mental life

In evaluating whether children's evaluations of mental life varied by target, we submitted our data to two mixed-effects linear models (one for psychological agency, and one for psychological experience). Children in our sample ascribed the greatest degree of psychological agency to the human ($M = 2.26$, $SD = 1.06$), followed by the robot ($M = 2.19$, $SD = 1.11$), and toy bear ($M = 0.90$, $SD = 1.24$), though children evaluated the agency of the human and robot targets similarly (human-robot: $\beta = -0.07$, 95% CI [−0.43, 0.28], $t(459) = -0.41$, $p = .68$). Children only rated the toy bear as having diminished psychological agency (human-toy bear: $\beta = -1.37$, 95% CI [−1.73,−1.00], $t(459) = -7.38$, $p < .001$, see the leftmost panel of Fig. 2).

For psychological experience, children granted the human with the highest degree of mental life ($M = 2.50$, SD = 0.92), which differed from the robot ($M = 1.75$, $SD = 1.32$), which differed from the toy bear ($M = 1.25$, $SD = 1.39$; human-robot: $\beta = -0.74$, 95% CI [−1.11,−0.37], $t(457) = -3.93$, $p < .001$; human-toy bear: $\beta = -1.24$; 95% CI [−1.63, −0.86], $t(457) = -6.38$, $p < .001$, see the second panel from the left in Fig. 2).

3.5. Harm vulnerability

When evaluating these targets' degree of suffering, the children in our sample distinguished between targets to some extent: The human target was rated the most harmed ($M = 2.78$, $SD = 0.66$), followed by the robot ($M = 2.40$, $SD = 1.00$) and the toy bear ($M = 2.04$, $SD = 1.27$)—though only the toy bear was rated significantly less harmed than the human target (human-robot: $\beta = -0.37$, 95% CI [−0.75, 0.008], $t(226) = -1.93$, $p = .06$; human-toy bear: $\beta = -0.73$, 95% CI [−1.13, −0.34], $t(226) = -3.68$, $p < .001$, see the middle panel of Fig. 2).

3.6. Agent punishment and intentionality

We also probed whether children thought the agent (i.e., the non-target entity who performed the transgressions against the target) deserved punishment for transgressing, as well as whether children thought this behavior was performed intentionally.

In the present sample, children endorsed that agents who transgressed against humans ($M = 2.58$, $SD = 0.69$) and robots ($M = 2.43$, $SD = 0.99$) deserved similar punishment ($\beta = -0.15$, 95% CI [−0.55, 0.25], $t(227) = -0.74$, $p = .46$), over and above that of the agent who transgressed against the toy bear ($M = 1.56$, $SD = 1.32$; $\beta = -1.02$, 95% CI [−1.44, −0.61], $t(227) = -4.85$, $p < .001$, see the rightmost panel of Fig. 2). They did not, however, endorse that that these agents differed on intentionality (human-robot: $\beta = -0.34$, 95% CI [−0.83, 0.15], $t(227) = -1.36$, $p = .18$; human-toy bear: $\beta = -0.29$, 95% CI [−0.80, 0.21], $t(227) = -1.14$, $p = .25$, see the second panel from the right Fig. 2).

From this experiment, we gained initial insight into children's evaluations of robot mental life and moral standing. Children rated the human and robot targets similarly on most metrics—such as higher-order mental life, vulnerability to harm, and degree of transgressor punishment—while also distinguishing between these targets in terms of experiential mental life. This provides preliminary evidence that children privilege robot moral standing, as compared to artifacts (e.g., toys).

4. Experiment 2

Are these patterns stable across development, such that adults think similarly about robots? For Experiment 2, we provided the same experiment to a sample of adults, addressing whether the patterns observed in Experiment 1 persist beyond childhood.

We preregistered two predictions: First, we anticipated that adults would rate the human target as having the most mental life and being the most vulnerable to harm—more so than either the robot or toy bear targets. Second, we predicted that adults would rate the robot as having

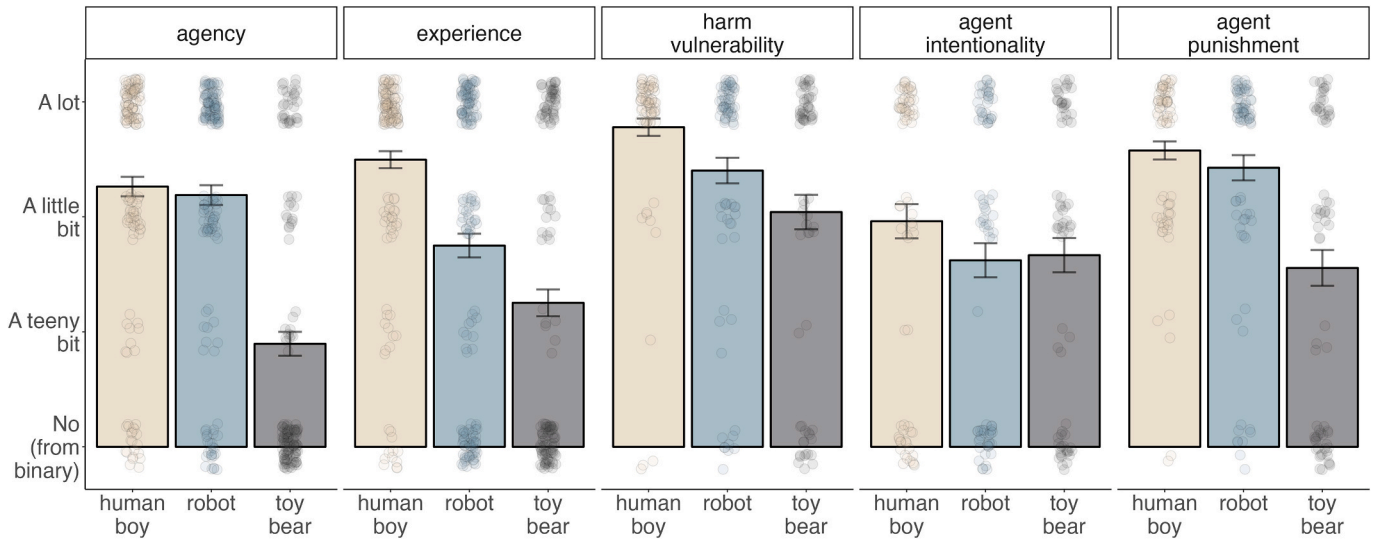

Fig. 2. Children's evaluations from Experiment 1. Error bars depict +/− standard error of the mean.

greater mental life (i.e., agency and experience together) than the toy, but they would rate the robot and toy as similarly vulnerable to harm.

4.1. Method

4.1.1. Participants

We recruited 158 online participants from Amazon Mechanical Turk (anticipating a sample of approximately 125, after exclusions). We excluded two participants for failing attention checks, leaving 156 in our final sample (M_{age} = 33.8, SD_{age} = 9.79; 78 identified as male, 71 identified as female, 1 identified as nonbinary, 6 declined to answer). To participate, we required participants to be located within the United States and have an approval rating of 95% or greater. We paid participants $0.25 for participation.

4.1.2. Materials and procedure

All materials were identical to those of Experiment 1 (with the addition of an adult-appropriate attention check).

4.2. Results

4.2.1. Mental life

To test our hypotheses, we ran a series of mixed-effects models.[2] As in Experiment 1, we re-coded the data to create a continuous measure of responses. Here, adults rated the human target highest in agency (M = 2.49, SD = 0.85), over and above the robot (M = 1.79, SD = 1.29; β = −0.70, 95% CI [−1.00, −0.40], $t(606) = -4.59$, $p < .001$) and toy bear (M = 0.35, SD = 0.80; β = −2.15, 95% CI [−2.44, −1.85], $t(606) = -14.14$, $p < .001$; see leftmost panel of Fig. 3).

For experience, however, we observed stronger distinctions in how adults evaluated human and non-human targets. The adults in our sample only endorsed the human target ($M = 2.63$, $SD = 0.82$) as having experiential mental life, and they generally denied these capacities for both the robot ($M = 0.34$, $SD = 0.86$) and toy bear ($M = 0.29$, $SD = 0.78$; see the second panel from the left in Fig. 3). This was convergent with results from a mixed-effects linear model with participant as a random effect (human-robot: β = −2.29, 95% CI [−2.56, −2.03], $t(605) = -16.89$, $p < .001$; human-toy bear: $\beta = -2.34$, 95% CI [−2.61, −2.07], $t(605) = -17.30$, $p < .001$).

We also predicted that, when looking at agency and experience together, adults would evaluate the robot as having greater mental life than the toy. These data suggested this to be the case, $t(650.13) = 9.72$, 95% CI [0.59, 0.89], $p < .001$.

4.2.2. Harm vulnerability

In contrast to the children tested in Experiment 1, adults demonstrated a distinct pattern in evaluating targets' susceptibility to harm (human-robot: $\beta = -1.51$, 95% CI [−1.91, −1.11], $t(300) = -7.42$, $p < .001$; human-toy bear: β = −1.40, 95% CI [−1.80, −1.00], $t(300) = -6.92$, $p < .001$; see middle panel of Fig. 3). As anticipated, the human was seen most capable of suffering (M = 2.27, SD = 0.84), without differences between the robot ($M = 0.76$, $SD = 1.23$) and toy bear targets (M = 0.86, SD = 1.27), $t(198.90) = -0.57$, 95% CI [−0.45, 0.25], p = .57. An exploratory model further suggested that participants' tendency to endorse suffering was strongly predicted by the denial of experience-related mental capacities ($\beta = 0.77$, 95% CI [0.62, 0.91], $t(150) = 10.15$, $p < .001$) but not agency-related capacities ($\beta = -0.14$, 95% CI [−0.30, 0.03], $t(150) = -1.64$, $p = .10$).

4.2.3. Agent punishment and intentionality

Moreover, adults also thought transgressors against robots (β = −0.37, 95% CI [−0.74, −0.003], $t(301) = -1.99$, $p < .05$) and toy bears ($\beta = -0.93$, 95% CI [−1.29,−0.56], $t(301) = -4.95$, $p < .001$) deserved less punishment than transgressors against humans (see the rightmost panel of Fig. 3). Similarly to children's evaluations, however, these effects seem unrelated to beliefs about transgressor intentionality ($ps \geq 0.05$; see second panel from the right in Fig. 3).

5. Experiment 3

The prior two experiments suggest a developmental distinction: In Experiment 1, children considered robots as having rich mental lives and moral standing (e.g., believing that they can suffer, endorsing punishment towards transgressors). Adults in Experiment 2, by contrast, typically denied these attributes. When do children begin to constrict their moral circle towards robots?

In a third experiment, we gauged children's beliefs about the moral

[2] We initially preregistered a series of ANOVAs for this experiment. Given the structure of our data, however, we later agreed that running linear mixed-effects models (with participant set as a random effect) would be superior for testing our hypotheses. As such, we report linear mixed-effects models (and binomial logistic regressions, in the SOM) throughout the paper. This is the only deviation from our preregistration within this experiment, and we report results from the original analysis plan in the SOM. These alternative models do not change any of the reported effects.

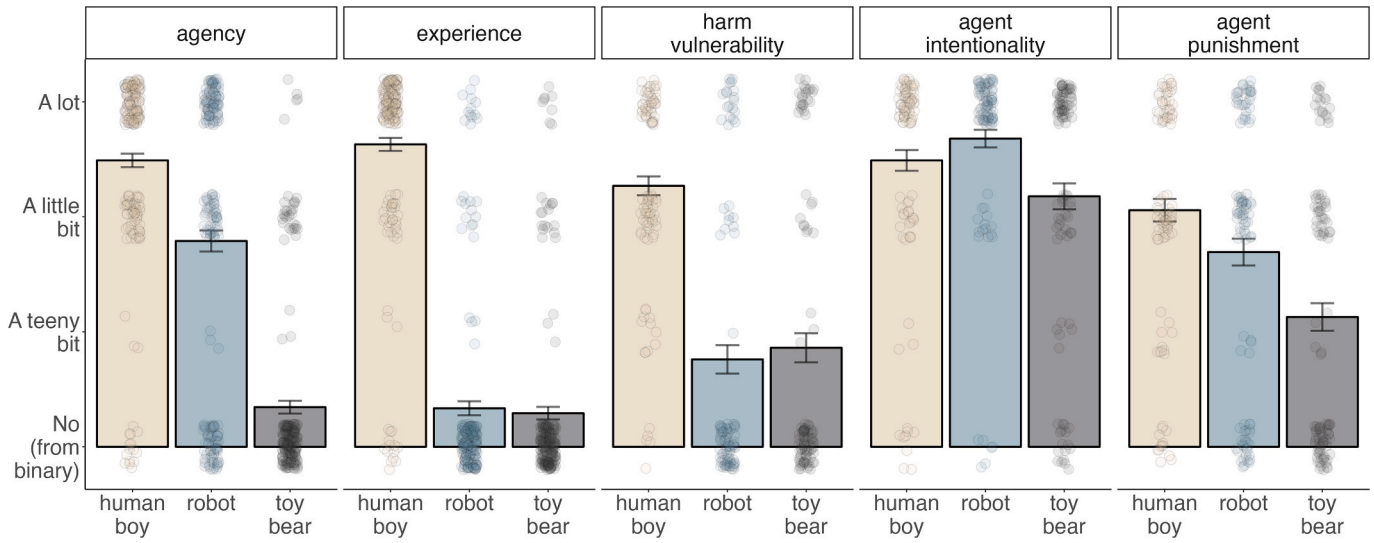

Fig. 3. Adults' evaluations from Experiment 2. Error bars depict +/− standard error of the mean.

standing of artificial intelligence by asking whether it was "okay" or "not okay" to transgress against a robot (as compared to the human and toy bear targets). We call these "moral permissibility" beliefs, as they regard whether it was "okay" that a transgression occurred. We see this as a more valid test of moral standing than our previous metric, as one might endorse that an entity can suffer without necessarily caring morally that the entity suffers.

We improved this experiment in two additional ways: First, given our inability to probe age effects in Experiment 1, we took care to sample enough children within each age group to identify potential age-related differences with sufficient power (95% for a three-way interaction between age, target, and mental capacity ascription). Second, we modified this experiment to be within-subjects. We believe this provides important insight into how children compare these targets to one another when forming judgments.

In light of this, we preregistered three predictions. We anticipated that (1) older children would most strongly endorse human moral standing, whereas younger children would discriminate between humans, robots, and toys to a lesser degree. We also predicted that older children would ascribe (2) heightened agency and (3) heightened experience to the human boy target (over the robot and toy bear), whereas younger children would discriminate between these targets to a lesser extent.

5.1. Method

5.1.1. Participants

In light of a G*Power-based power analysis, we collected data from 161 children (ages 5–10.99). This allowed us to test for a medium-sized ($f = 0.25$) three-way interaction with 95% power. Before analysis, we removed data from 13 children (due to failing embedded manipulation checks, being outside the age boundary for the study, or for not having age-related demographic information). Our final sample consisted of 148 children ($M_{age} = 7.59$, $SD_{age} = 1.64$; 73 identified as male, 75 identified as female). Each child received a small prize for participating.

5.2. Materials and procedure

As in Experiment 1, we provided participants with vignettes where a boy transgressed against either another human boy, a robot, or a toy bear. We removed the harm vulnerability item and replaced it with an updated moral permissibility item (e.g., "Do you think it was **okay** for Matt to push Drew onto the ground, or that it was **not okay** for Matt to push Drew onto the ground?"). We followed up the initial binary items with an extended three-point scale, as described in Experiment 1.

In light of this experiment being entirely within-subjects and provided to children, we opted to only include the two most representative mental capacity items for agency and experience, respectively (i.e., the capacity to tell right from wrong, to act with self-control, to experience hunger, and to experience fear). Further, given that children (Experiment 1) and adults (Experiment 2) ascribed similar degrees of intentionality to transgressors, we removed the item probing whether the agent "meant to" transgress. Aligning with our prior methodology, however, we again asked participants whether the transgressor deserved punishment. We also collected data from children online (via Zoom conference call, with Qualtrics stimuli screen-shared to participants) or in-person using iPads (via the iOS Qualtrics app).[3]

5.3. Results

5.3.1. Mental life

We first analyzed a mixed-effects linear model (setting the human target as the reference category and participant as a random effect) to examine whether children's ascription of mental life varied by target. Consistently with the prior experiments, children distinguished between the mental lives of each target (see the top two panels of Fig. 4). As compared to the human, participants ascribed less mental life to the robot ($\beta = -0.97$, 95% CI [−1.10, −0.84], $t(1771) = -15.10$, $p < .001$) and the toy bear ($\beta = -1.77$, 95% CI [−1.89, −1.64], $t(1771) = -27.51$, $p < .001$)—an effect which amplified with age (human-robot: $\beta = -0.28$, 95% CI [−0.36, −0.21], $t(1768) = -7.41$, $p < .001$; human-toy: $\beta = -0.30$, 95% CI [−0.37, −0.22], $t(1768) = -7.80$, $p < .001$). Indeed, children seemed to learn to deny psychological experience to robots, in particular. Though, with age, children ascribed diminished agency to both the robot ($\beta = -0.22$, 95% CI [−0.33, −0.12], $t(880) = -4.20$, $p < .001$) and the toy ($\beta = -0.35$, 95% CI [−0.46, −0.25], $t(880) = -6.63$, $p < .001$), this age-related dampening for robots was especially strong for psychological experience ($\beta = -0.34$, 95% CI [−0.45, −0.24], $t(880) = -6.55$, $p < .001$).

[3] We began collecting data for this experiment prior to the onset of the COVID-19 pandemic. Our lab collected the remainder of the data via Zoom. We made a slight modification to the phrasing of the experiment—using the word "cannot," rather than "can't"—for ease of understanding participant choices over the computer. These shifts in phrasing did not impact the results of the experiment.

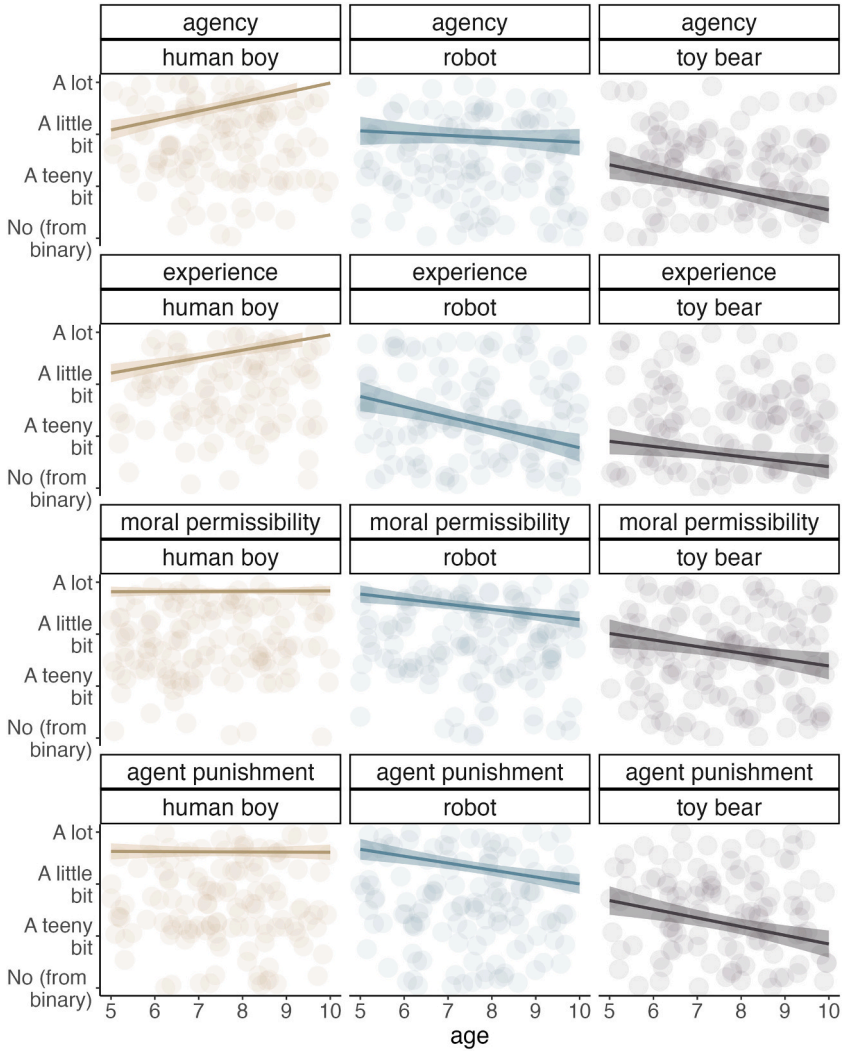

Fig. 4. Scatterplot of children's evaluations of agency (top panel), experience (second panel from the top), moral permissibility (second panel from the bottom), and punishment towards transgressors (bottom panel) from Experiment 3 (with 95% confidence intervals). The x-axis denotes participant age.

Denying robots certain psychological capacities may underpin an emerging lack of moral concern towards them. Though children's evaluations of psychological agency did predict the degree to which it was "okay" to transgress against a given target ($\beta = 0.13$, 95% CI [0.05, 0.20], $t(437) = 3.15$, $p < .003$), children's evaluations of psychological experience was a stronger predictor than this metric ($\beta = 0.24$, 95% CI [0.15, 0.32], $t(437) = 5.56$, $p < .001$).

5.3.2. *Moral permissibility*

We also fit a model to examine whether permissibility beliefs varied by target. Collapsing across age, children saw transgressions as "more okay" when directed towards the robot ($\beta = -0.31$, 95% CI [−0.43, −0.18], $t(883) = -4.80$, $p < .001$) and toy bear ($\beta = -1.14$, 95% CI [−1.26, −1.01], $t(883) = -17.73$, $p < .001$). We then added participant age into this model. With age, children rated transgressions against robots ($\beta = -0.10$, 95% CI [−0.18, −0.03], $t(880) = -2.62$, $p < .01$) and toys ($\beta = -0.13$, 95% CI [−0.20, −0.05], $t(880) = -3.30$, $p < .002$) as more permissible than transgressions against the human (see the second panel from the bottom in Fig. 4).

5.3.3. *Agent punishment*

Finally, we fit an exploratory mixed-effects linear model predicting children's punishment judgments. Here, older children tended to endorse lesser punishment for transgressors against robots and toys (human-robot: $\beta = -0.13$, 95% CI [−0.21, −0.05], $t(879) = -3.02$, $p < .004$; human-toy: $\beta = -0.16$, 95% CI [−0.25, −0.08], $t(879) = -3.82$, $p < .001$).

Taken together, these findings fit well with the possibility that, early in development, children are willing to grant moral standing to robots. With age, we observed a more adult-like pattern of judgments: Children ascribed less mental life to robots, as compared to humans—with this pattern being especially pronounced for psychological experience. Further, when children endorsed that robots had less psychological experience, they also tended to believe that it was "more okay" to transgress against them. This dovetailed with their beliefs about punishment (i.e., that it was more acceptable to transgress, and that transgressors deserved lesser punishment). We take this to suggest that children may learn over early childhood to think of robots as less worthy of moral consideration.

6. Experiment 4

Data from the prior experiments demonstrated a developmental shift in people's beliefs concerning robot mental life and moral standing. In a final experiment, we provided children with a behavioral measure of moral standing—a Dictator Game. We did this in an attempt to approximate a high-cost measure of moral concern. Past work has shown that children's moral reasoning is linked to high-cost, but not low-cost

prosocial behavior (Eisenberg & Shell, 1986), with cost being a major barrier to prosocial action (Kirby et al., 2023; Mao et al., 2023). Although perhaps debatable from a philosophical standpoint, we see a meaningful distinction between someone who *claims* to care about climate change (but never actually *tries* to reduce their carbon footprint) versus an individual who claims to care (and actually tries to reduce their carbon footprint). As such, we provide a costly measure of prosociality as a more stringent test of children's moral concern. By allowing participants to distribute valuable resources to these targets, we gain deeper understanding of whether children truly consider robots to be a social and moral other.

We preregistered a series of developmental predictions.[4] Specifically, we anticipated that (1) children would share the most tokens during the Dictator Game with the human target, followed by the robot target, and then the control targets. We also predicted that (2) older children would donate fewer tokens to all non-human targets than younger children, and (3) perceived mental life (collapsing across agency and experience) would predict greater donation across all targets.

6.1. Method

6.1.1. Participants

To detect a small to medium-sized simple effect ($d = 0.25$) with 85% power, we needed to obtain data from 150 children (with 25 per age group, ages 5–10.99). Our final sample included 151 children (M_{age} = 7.49, SD_{age} = 1.71; 76 identified as male, 74 identified as female, 1 identified as "other"). We did not exclude any participants. Each child received a small prize for their participation.

6.1.2. Materials and procedure

As with Experiment 3, we collected data from children either online or in-person. For this experiment, we included the three prior targets, as well as a "rock" (control). Also identically to Experiment 3, we prompted children to evaluate the mental life of each target (within-subjects) on four mental life items.

In a separate phase of the experiment, we had children partake in a Dictator Game (Benenson et al., 2007) with each target. We described the situation as follows: "Here, we have four tokens. You can keep as many as you want or give away as many as you want. If you give a token away, it will go to the [target]. You can use the tokens that you keep to pick out a prize at the end. What do you want to do with your tokens?". Children indicated to the experimenter how many tokens they wished to keep or distribute.

6.2. Results

6.2.1. Mental life

To determine whether children's beliefs about higher-order mental life varied by target and by participant age, we ran mixed-effects linear models with participant set as a random effect (see Fig. 5). Consistently with the prior findings, older children ascribed diminished agency to the robot ($\beta = -0.12$, 95% CI [−0.21, −0.02], $t(1196) = -2.38$, $p < .02$), and even less to the toy bear ($\beta = -0.21$, 95% CI [−0.30, −0.11], $t(1196) = -4.28$, $p < .001$) and rock ($\beta = -0.19$, 95% CI [−0.29, −0.10], $t(1196) = -3.92$, $p < .001$).

Children distinguished between the human and non-human targets to an even greater degree for evaluations of psychological experience. We ran the same model on ratings of experiential mental life: Again, children ascribed less experience, as age increased, to the robot ($\beta = -0.32$, 95% CI [−0.40, −0.25], $t(1195) = -8.52$, p < .001), toy bear ($\beta = -0.21$, 95% CI [−0.28, −0.13], $t(1195) = -5.46$, $p < .001$), and rock ($\beta = -0.18$, 95% CI [−0.25, −0.10], $t(1195) = -4.72$, $p < .001$).

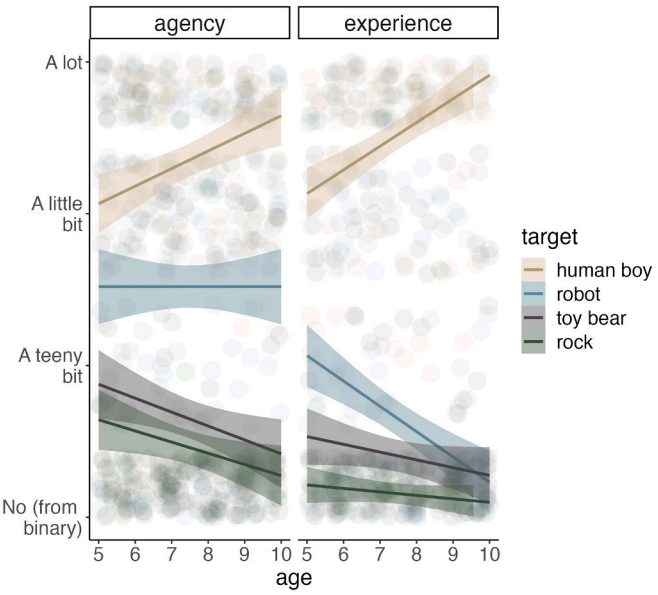

Fig. 5. Scatterplot of children's evaluations of mental life from Experiment 4 (with 95% confidence intervals). The x-axis denotes participant age.

6.2.2. Dictator Game

We then tested whether children donated resources differently towards each of the targets (see Fig. 6), and whether donation behavior shifted with age. Contrary to our predictions, after Bonferroni correction, children's donation behavior only significantly differed between the human and rock ($p < .001$), and the robot and rock ($p < .03$). To further probe potential mechanisms underpinning children's moral behavior, we followed this analysis with a mixed-effects model with participant set as a random effect: With age, children donated fewer tokens to the robot ($\beta = -0.12$, 95% CI [−0.23, −0.001], $t(594) =$

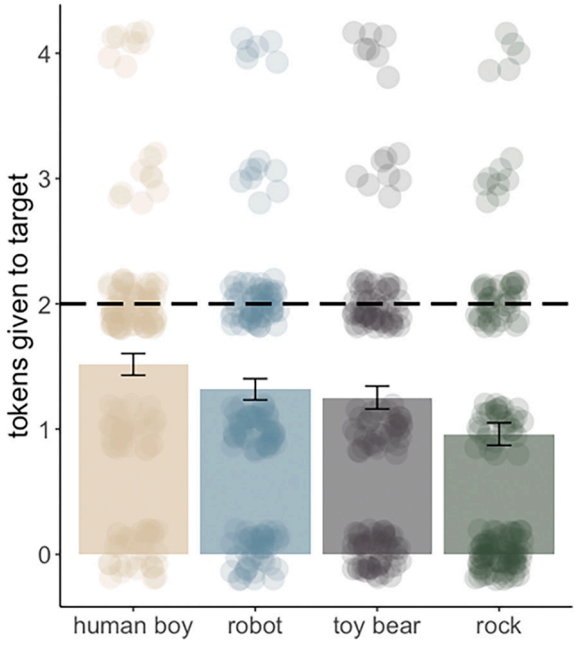

Fig. 6. Children's average donation behavior by target in Experiment 4. The dashed line indicates a "fair" donation in the Dictator Game (2 out of 4 tokens). Error bars depict +/− standard error of the mean.

[4] We originally planned to replicate this study with adults, and we preregistered developmental-adult comparisons. Due to constraints, we focus here solely on predictions concerning the developmental sample. We see it as valuable to execute an extension of this paradigm with adults in future research.

−2.00, $p < .047$), the toy bear ($\beta = -0.13$, 95% CI [−0.24, −0.01], $t(594) = -2.16$, $p < .033$), and the rock ($\beta = -0.21$, 95% CI [−0.32, −0.09], $t(594) = -3.53$, $p < .001$), as compared to the human target.

Following this, we ran an additional model including mental life judgments as a predictor. Here, we find that, even when adjusting for target, children's ascriptions of mental life predicted donation behavior (see Fig. 7; agency: $\beta = 0.11$, 95% CI [0.008, 0.21], $t(596) = 2.14$, $p < .034$; experience: $\beta = 0.12$, 95% CI [0.003, 0.24], $t(596) = 2.02$, $p < .045$).

In sum, these data replicate and extend the patterns observed previously. Over early to middle childhood, children increasingly denied mental life to robots. When children endorsed robots (and other non-human targets) as having lesser mental life, they also tended to donate fewer tokens to these targets in a Dictator Game. Importantly, children of all ages consistently ascribed mental life to the human target, and they most often donated a "fair" number of tokens during human trials. By contrast, during robot trials, we observed substantial variability in perceived mental life—which, again, correlated with the number of tokens donated.

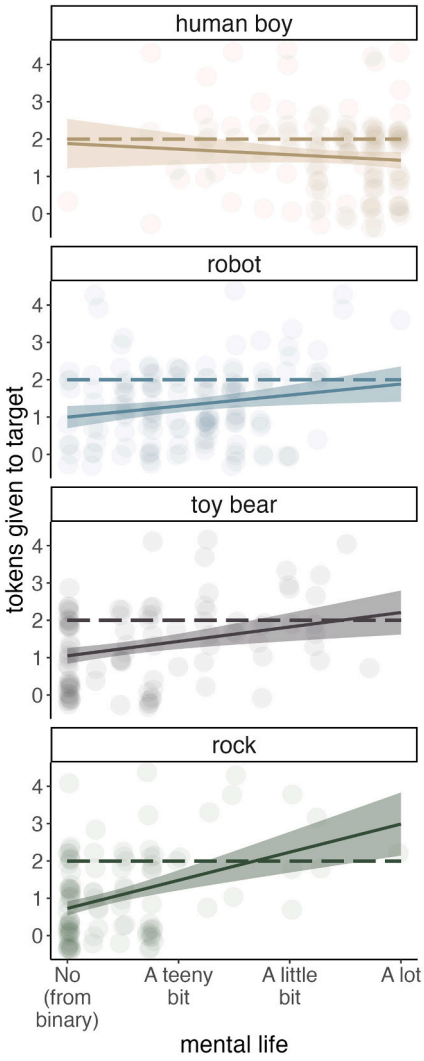

Fig. 7. Scatterplot comparing children's donation behavior and evaluations of mental life (collapsed across agency and experience) from Experiment 4 (with 95% confidence intervals). The dashed line indicates a "fair" donation in the Dictator Game (2 out of 4 tokens).

7. General discussion

Robots—and artificial intelligence, more broadly—are becoming increasingly present in everyday life. In four experiments, we examined how children and adults consider the mental life and moral standing of robots, tracing the developmental trajectory of their beliefs.

In Experiment 1, we observed that children tended to consider robots' "higher-order" mental life as similar to humans, while also denying that robots had experiential mental capacities. They also thought that, following a moral transgression, humans and robots suffered to a similar degree—over and above the suffering experienced by a toy bear (control). They also believed that transgressors against humans and robots deserved similar punishment, whereas a transgressor against a toy bear deserved less.

In Experiment 2, we replicated this experiment with a sample of adults. Here, we observed a different set of patterns: Adults thought robots had less agentic and experiential mental life than humans, far closer to their evaluations of the toy. They also indicated that the robot suffered less than the human after being transgressed against, which coincided with their denial of the robot having an experiential mind. Unlike the children in Experiment 1, adults further believed that transgressors against robots deserved less punishment than transgressors against humans.

In Experiments 3 and 4, we gained insight into the developmental trajectory of children's beliefs. In Experiment 3, we modified our moral standing measure to gauge whether children thought transgressions against robots were "more okay" than transgressions against humans. Indeed, with age, children endorsed that transgressions against robots were more morally permissible than transgressions against humans—which was predicted by their greater tendency to deny that robots had experiential mental life. We also observed that, with age, children tended to ascribe lesser agentic mental capacities to robots (as compared to the human target). Finally, children most often said that transgressors against humans deserved punishment for their wrongdoing—and with age, they endorsed lesser punishment for transgressors against robots and toys.

In Experiment 4, we probed children's moral behavior towards robots. We again tested their evaluations of mental life and observed whether this predicted donation behavior in a Dictator Game. Aligning with the prior experiments, we observed a developmental decline in children's thinking that robots had agentic and experiential mental capacities. This decline predicted their donation behavior: Older children donated fewer tokens and ascribed fewer mental capacities to robots (along with both control targets, a toy bear and a rock).

Taken together, this suggests that the tendency to deny the moral standing of robots may emerge over the course of early development.

7.1. Do children really see robots as having mental life and moral standing?

One unexpected finding emerged regarding younger children's relatively high attribution of mental life and moral standing to the control targets, the toy bear and the rock. This raises the concern that the participants were pretending or play-acting when responding to our vignettes. Indeed, it may be that "children who more readily imagine (or simulate) others' mental states do so both in the context of role play *and* anthropomorphism" (Severson & Woodard, 2018, p. 11, emphasis added). This concern was part of the motivation for Experiment 4, though the findings did not completely alleviate these concerns. If participants were merely pretending that the rock had moral standing, they shouldn't have given any resources to it—and yet, some participants did.

This complicates the interpretation of our data. Were the younger children "playing around" when judging the robot target's mental life and moral standing? Given existing evidence that children maintain more expansive moral concern early in development (e.g., Neldner et al., 2018; Wilks et al., 2021), which coincides with a documented tendency

for younger children to ascribe robots with mental life (e.g., Kahn et al., 2012), we remain open to the possibility that the younger children we tested considered robots to have true mental life and moral standing. We note, however, that donation behavior may have been artificially inflated by our paradigm: Children may have been more generous than they would have otherwise been, as they made all of their donation decisions in front of an experimenter.

Relatedly, there are also questions about the relationship between perceived moral standing and donation behavior. Past work has shown that several factors affect children and adults' willingness to donate resources to robots (as compared to humans), including perceptions of the robot (e.g., likability, anthropomorphism, and utility; de Kleijn et al., 2019), social status (Weiß et al., 2020), and prior success in multi-round games (Hsieh et al., 2023). As such, though we employed this task to approximate a high-cost moral belief, we recognize that a number of additional factors may influence children's decision to share (or not share) in this case. Importantly, however, the relationship identified between perceived mind and sharing indicates that the measure is at least somewhat related to perceived moral status. Nonetheless, we recommend that future research probe the present effect further, potentially by allowing children to engage in private donations or by employing alternative behavioral measures.

7.2. Limitations

We also recognize that a series of methodological and theoretical limitations apply to these experiments. First, we focused on a narrow set of static stimuli and collected data exclusively from the United States and Canada. Though these sample demographics may limit whether the present effects generalize (Simons et al., 2017), recent evidence suggests that researchers' study design and analysis choices may prove an even greater threat to generalizability (Holzmeister et al., 2024; Yarkoni, 2022). Testing the robustness of the present effects by, for example, modifying the existing paradigm to gauge children's moral judgments and behavior towards an *actual,* physically present robot (e.g., Mollahosseini et al., 2018) or by taking a "multi-analyst" approach (Aczel et al., 2021) may be particularly informative.

Second, we cannot speak to whether children interpreted our "moral permissibility" item (i.e., whether it was "okay" to transgress) as regarding *intrinsic* or *extrinsic* value. We intended to ask children about intrinsic value—when an entity matters morally for its own sake. But when something has extrinsic value, it is good "for the sake of something else to which it is related in some way" (Zimmerman & Bradley, 2019). If we had given children the opportunity to provide open-ended responses, we would have better insight into whether they interpreted our item as intended. This also may serve as a valuable avenue for future study.

7.3. Future directions

We see the current set of experiments as a foray into understanding developmental moral standing beliefs, particularly within the domain of artificial intelligence. Many future directions remain. Children may ascribe greater moral standing to robots, but the implications of these beliefs are unclear. What happens, for instance, if protecting a robot is in tension with protecting a human? Some existing literature gestures at how children might evaluate these cases (Kahn et al., 2012), but there remains ample opportunity for further research to shed light on this.

This work also raises questions concerning mechanism. Here, we highlight the relationship between mental life and moral standing, echoing an expansive existing literature in moral psychology (e.g., Gray et al., 2007; Schein & Gray, 2018). There are a host of other possible moderators which may contribute to perceptions of moral standing. One such possibility is perceived dangerousness (Piazza et al., 2014). It may be that adults' apprehension towards granting AI with moral privileges stems in part from perceiving them as harmful to human interests.

We note as well that these findings may be limited to children's beliefs about robots and not to artificial intelligence more generally. Critically, the "robot" we asked children about had a humanoid physical form—unlike the chatbots, for example, that are becoming increasingly familiar to children and adults alike (e.g., ChatGPT). How children evaluate the mental capacities and moral standing of AI will likely vary across specific kinds of technology (Flanagan et al., 2023; Ladak et al., 2023). In this same vein, we see it as valuable to probe potential individual differences that could moderate these effects, such as familiarity with common AI products (e.g., voice assistants, like Apple's Siri) or endorsement of "substratism" (i.e., having prejudicial attitudes towards AI as a result of their non-biological material composition; Pauketat & Anthis, 2022).

In sum, we report evidence from four studies demonstrating a developmental shift in children's evaluations of robot moral standing and mental life. Earlier in childhood, robots appear conceptually closer to natural kinds, like humans, than artifacts, like toys—both within the mental and the moral domain. This sheds new light on children's moral cognitive development and their relationship to emerging technologies.

CRediT authorship contribution statement

Madeline G. Reinecke: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matti Wilks:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Paul Bloom:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Data curation, Conceptualization.

Acknowledgements

We would like to thank the Jacobs Foundation for their support of this research (Award to PB: 70000890). MGR is supported by the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (Award: NIHR203316) and the Wellcome Trust (Award: Wellcome Centre for Ethics and Humanities, 203132/Z/16/Z). The views expressed are those of the authors and do not necessarily reflect those of the funding bodies or by the Department of Health and Social Care. We also extend our gratitude to the members of the Mind and Development Lab (Yale University) and the Centre for Mind and Morality (University of Toronto), as well as our team of research assistants.

Appendix A. Supplementary data

The Supplementary Online Materials (SOM) for this article can be accessed at https://doi.org/10.1016/j.cognition.2024.105983 or on the Open Science Framework.

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., … Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife, 10*, Article e72185. https://doi.org/10.7554/eLife.72185

Benenson, J. F., Pascoe, J., & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior, 28*(3), 168–175. https://doi.org/10.1016/j.evolhumbehav.2006.10.003

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny Valley feelings are acquired in childhood. *Child Development, 90*(4), 1202–1214. https://doi.org/10.1111/cdev.12999

Caviola, L., Schubert, S., Kahane, G., & Faber, N. S. (2022). Humans first: Why people value animals less than humans. *Cognition, 225*, Article 105139. https://doi.org/10.1016/j.cognition.2022.105139

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., & Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports, 12*(1), 4560. https://doi.org/10.1038/s41598-022-08460-1

Eisenberg, N., & Shell, R. (1986). Prosocial moral judgment and behavior in children: The mediating role of cost. *Personality and Social Psychology Bulletin, 12*(4), 426–433. https://doi.org/10.1177/0146167286124005

Flanagan, T., Wong, G., & Kushnir, T. (2023). The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Developmental Psychology, 59*(6), 1017–1031. https://doi.org/10.1037/dev0001524

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. https://doi.org/10.1126/science.1134475

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2024). Heterogeneity in effect size estimates. *Proceedings of the National Academy of Sciences, 121*(32), Article e2403490121. https://doi.org/10.1073/pnas.2403490121

Hsieh, T.-Y., Chaudhury, B., & Cross, E. S. (2023). Human–Robot Cooperation in Economic Games: People Show Strong Reciprocity but Conditional Prosociality Toward Robots. *International Journal of Social Robotics, 15*(5), 791–805. https://doi.org/10.1007/s12369-023-00981-7

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences, 111*(35), 12710–12715. https://doi.org/10.1073/pnas.1402280111

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., … Shen, S. (2012). "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental Psychology, 48*(2), 303–314. https://doi.org/10.1037/a0027033

Kirby, J. N., Kirkland, K., Wilks, M., Green, M., Tanjitpiyanond, P., Chowdhury, N., & Nielsen, M. (2023). Testing the bounds of compassion in young children. *Royal Society Open Science, 10*(2), Article 221448. https://doi.org/10.1098/rsos.221448

de Kleijn, R., van Es, L., Kachergis, G., & Hommel, B. (2019). Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human-Computer Studies, 122*, 168–173. https://doi.org/10.1016/j.ijhcs.2018.09.008

Ladak, A. (2024). What would qualify an artificial intelligence for moral standing? *AI and Ethics, 4*(2), 213–228. https://doi.org/10.1007/s43681-023-00260-1

Ladak, A., Wilks, M., & Anthis, J. R. (2023). Extending perspective taking to nonhuman animals and artificial entities. *Social Cognition, 41*(3), 274–302. https://doi.org/10.1521/soco.2023.41.3.274

Manzi, F., Peretti, G., Di Dio, C., Cangelosi, A., Itakura, S., Kanda, T., Ishiguro, H., Massaro, D., & Marchetti, A. (2020). A robot is not worth another: Exploring Children's mental state attribution to different humanoid robots. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.02011

Mao, Y., Reinecke, M. G., Kunesch, M., Duéñez-Guzmán, E. A., Comanescu, R., Haas, J., & Leibo, J. Z. (2023). Doing the right thing for the right reason: Evaluating artificial moral cognition by probing cost insensitivity. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2305.18269

Martin, D. U., Perry, C., MacIntyre, M. I., Varcoe, L., Pedell, S., & Kaufman, J. (2020). Investigating the nature of children's altruism using a social humanoid robot. *Computers in Human Behavior, 104*, Article 106149. https://doi.org/10.1016/j.chb.2019.09.025

Mollahosseini, A., Abdollahi, H., Sweeny, T. D., Cole, R., & Mahoor, M. H. (2018). Role of embodiment and presence in human perception of robots' facial cues. *International Journal of Human-Computer Studies, 116*, 25–39. https://doi.org/10.1016/j.ijhcs.2018.04.005

Neldner, K., Crimston, C., Wilks, M., Redshaw, J., & Nielsen, M. (2018). The developmental origins of moral concern: An examination of moral boundary decision making throughout childhood. *PLoS One, 13*(5), Article e0197819. https://doi.org/10.1371/journal.pone.0197819

Nijssen, S. R. R., Müller, B. C. N., van Baaren, R. B., & Paulus, M. (2019). Saving the robot or the human? Robots who feel deserve moral care. *Social Cognition, 37*(1), 41–56. https://doi.org/10.1521/soco.2019.37.1.41

Paruzel-Czachura, M., Maier, M., Warmuz, R., Wilks, M., & Caviola, L. (2024). Children value animals more than adults do: A conceptual replication and extension. *Personality and Social Psychology Bulletin, 01461672231219391*. https://doi.org/10.1177/01461672231219391

Parviainen, J., & Coeckelbergh, M. (2021). The political choreography of the Sophia robot: Beyond robot rights and citizenship to political performances for the social robotics market. *AI & SOCIETY, 36*(3), 715–724. https://doi.org/10.1007/s00146-020-01104-w

Pauketat, J. V. T., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior, 136*, Article 107372. https://doi.org/10.1016/j.chb.2022.107372

Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition, 131*(1), 108–124. https://doi.org/10.1016/j.cognition.2013.12.013

Robotics, H. (2023). Sophia. In *Hanson Robotics*. https://www.hansonrobotics.com/sophia/.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Severson, R. L., & Woodard, S. R. (2018). Imagining Others' minds: The positive relation between Children's role play and anthropomorphism. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.02140

Shevlin, H. (2021). How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics, 30*(3), 459–471. https://doi.org/10.1017/S0963180120001012

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Sommer, K., Nielsen, M., Draheim, M., Redshaw, J., Vanman, E. J., & Wilks, M. (2019). Children's perceptions of the moral worth of live agents, robots, and inanimate objects. *Journal of Experimental Child Psychology, 187*, Article 104656. https://doi.org/10.1016/j.jecp.2019.06.009

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post*. https://www.washingtonpost.com/podcasts/post-reports/the-google-engineer-who-thinks-its-ai-has-come-alive/.

Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences, 13*(9), 397–402. https://doi.org/10.1016/j.tics.2009.06.008

NBC4 Washington. (2023). *AI robot panel tells UN press conference they could be better world leaders than humans*. https://www.nbcwashington.com/news/national-international/ai-robot-panel-tells-un-press-conference-they-could-be-better-world-leaders-than-humans/3380816/.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Children's intuitions about the structure of mental life. *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/5vj6j1f6.

Weiß, M., Rodrigues, J., Paelecke, M., & Hewig, J. (2020). We, them, and it: Dictator game offers depend on hierarchical social status, artificial intelligence, and social dominance. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.541756

Wilks, M., Caviola, L., Kahane, G., & Bloom, P. (2021). Children prioritize humans over animals less than adults do. *Psychological Science, 32*(1), 27–38. https://doi.org/10.1177/0956797620960398

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Zimmerman, M. J., & Bradley, B. (2019). Intrinsic vs. Extrinsic Value. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2019/entries/value-intrinsic-extrinsic/.