

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Developmental changes in perceived moral standing of robots

Permalink

<https://escholarship.org/uc/item/8f32d068>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN

1069-7977

Authors

Reinecke, Madeline G

Wilks, Matti

Bloom, Paul

Publication Date

2021

Peer reviewed

Developmental changes in perceived moral standing of robots

Madeline G. Reinecke (madeline.reinecke@yale.edu)

Department of Psychology, Yale University

Matti Wilks (matti.wilks@yale.edu)

Department of Psychology, Yale University

Paul Bloom (paul.bloom@yale.edu)

Department of Psychology, Yale University

Abstract

We live in an age where robots are increasingly present in the social and moral world. Here, we explore how children and adults think about the mental lives and moral standing of robots. In Experiment 1 ($N = 116$), we found that children granted humans and robots with more mental life and vulnerability to harm than an anthropomorphized control (i.e., a toy bear). In Experiment 2 ($N = 157$), we found that, relative to children, adults ascribed less mental life and vulnerability to harm to robots. In Experiment 3 ($N = 152$), we modified our experiment to be within-subjects and measured beliefs concerning moral standing. Though younger children again appeared willing to assign mental capacities — particularly those related to experience (e.g., being capable of experiencing hunger) — to robots, older children and adults did so to a lesser degree. This diminished attribution of mental life tracked with diminished ratings of robot moral standing. This informs ongoing debates concerning emerging attitudes about artificial life.

Keywords: morality; artificial intelligence; developmental psychology; mind perception

Introduction

In the show “Westworld,” viewers watched visitors to a “robot theme park” degrade, torture, and kill humanoids — robots who were virtually indistinguishable from the human visitors themselves. These robots seemed to have beliefs, desires, and bodily sensations, making it unclear which features (if any) drive potential differences in their moral worth (Bloom & Harris, 2018). Is killing a humanoid really equivalent to murder? One possibility is that category membership underpins moral standing judgments. Sophisticated as these robots may be, they will never count as true humans. Some term this perspective “speciesism.” Humans, by definition, stand atop the moral hierarchy (Singer, 2009).

Though artificial minds in the real world seem a far cry from human-level intelligence, these rapidly developing technologies raise important questions about moral standing (Risse, 2019). This paper sets aside normative matters, such as whether robots *actually* have moral standing (or whether they will in the future). We instead take up the descriptive claim: What do people think about the moral standing of artificial intelligence, and how do these beliefs form?

We approach this question 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 we are predisposed to favor those similar to us (Jordan, McAuliffe, & Warneken, 2014), we may similarly be predisposed to privilege the moral standing of humans over non-humans. Alternatively, children and adults may disagree about the moral standing of robots. 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 life 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 life may forecast the nature of human-robot socio-moral interaction.

Moral standing

Measures of speciesism appear to capture at least some aspects of moral judgment and behavior. People who tend to demonstrate “speciesist” attitudes (e.g., endorsing that “Morally, animals always count less than humans”) donate less time and money to charities related to animal well-being than those who rank lower in speciesism (Caviola, Everett, & Faber, 2019). Speciesism may similarly predict other real-life outcomes, such as meat-eating preferences (Caviola et al., 2019). Though the majority of this literature focuses on distinguishing humans versus non-human animals, these mechanisms may similarly apply to considering the moral worth of robots and artificial minds (Nijssen, Müller, van Baaren & Paulus, 2019). People may place artificial minds, as with many non-human animals, in the outer ranks of the moral circle.

These speciesist moral beliefs may emerge over time. Children seem to prioritize saving humans (at the expense of non-human animals) less often than adults do (Wilks, Caviola, Kahane, & Bloom, 2021), along with caring less about the moral concerns of (at least some) robots as they age (Sommer et al., 2019). This may result from mechanisms related to mind perception: Young children appear more willing than adults to ascribe mental abilities to artificial minds (Brink, Gray, & Wellman, 2019; Kahn et al., 2012; Weisman, Dweck, & Markman, 2017), which may correspond with their beliefs about moral standing (Gray, Gray, & Wegner, 2007). After witnessing an experimenter transgressing against a robot during a lab visit, for example, fifteen-year-olds proved less likely to see the robot as a “mental and moral other” than

nine-year-olds and twelve-year-olds (Kahn et al., 2012).

Our beliefs about artificial life may tie in with representing robots as part of a “new ontological category” (Kahn et al., 2011). Informed by developmental research (Kahn et al., 2012; Melson et al., 2009), the new ontological category hypothesis proposes that robots are not easily categorized as natural kinds or artifacts. Both children and adults may recognize them as having some, but not all, aspects of mental life (Jipson & Gelman, 2007; Nigam & Klahr, 2000). This coincides with research on speciesism. Placing robots in an “intermediary” category is consistent with granting robots diminished moral standing, relative to humans. Critically, however, the “new ontological category” hypothesis makes a further claim: If robots receive intermediate moral standing (e.g., Sommer et al., 2019), they must have greater moral standing than artifacts.

Taken together, we see this evidence as supporting a potential developmental shift in people’s beliefs about the mental life and moral standing of artificial minds: Though children and adults alike may distinguish robots from natural kinds and artifacts (Kahn et al., 2011; Gray & Wegner, 2012), this ontological gap may diminish with age (Jipson & Gelman, 2007; Wilks et al., 2021). In comparison to children, adults may represent robots as mentally and morally closer to artifacts than natural kinds.

We take up these possibilities in the present paper. In three experiments, we examine children and adults’ beliefs about the moral standing of a robot, human, and toy bear (control), as well as beliefs about the mental abilities of these targets. 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.

The preregistrations (when applicable), materials, analysis scripts, and data for all experiments are available on [ResearchBox](#).

Experiment 1

In our first experiment, we provided children with stories about a boy either transgressing against another human, transgressing against a robot, or transgressing against a toy bear. In one phase, participants evaluated the extent to which the target was harmed by these transgressions. In another phase, participants evaluated the mental capacities of their target.

We chose these three targets as stimuli to tease apart possible moral distinctions between artificial intelligence (“robot” condition) and humans (“human boy” condition). The toy bear served as a control, allowing us to identify whether children mentalize (and moralize) robots similarly to the artifacts commonly anthropomorphized during early development (e.g., toys) and also directly informs the ongoing discussion around the ontological categorization of artificial life.

Method

Participants. We collected data from 123 children between the ages of 4 and 13. We treated this as an exploratory exper-

iment, collecting data from participants recruited for a separate study. With this in mind, 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). Each child received a small prize at the end of the study for participating.

Materials and procedure. We randomly assigned participants to either the human boy, robot, or toy bear conditions (between-subjects). 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]”).

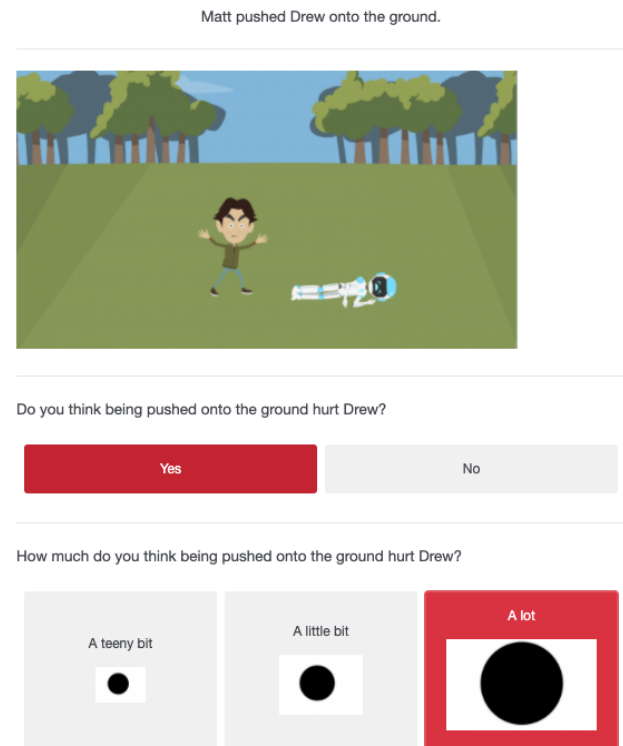


Figure 1: Example stimuli (physical transgression, robot condition) from Experiments 1 and 2.

In the “transgression” phase of the experiment, each participant evaluated one physical transgression and one non-physical transgression (in counterbalanced order) directed toward their target (see Figure 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 actor’s intentionality (i.e., “Do you think Matt meant [to push Drew onto

the ground]?”) and whether the actor deserved punishment (i.e., “Do you think Matt should get in trouble for [pushing Drew onto the ground]?”). We also prompted participants for beliefs about the perpetrator (e.g., whether the perpetrator intended to transgress). For brevity’s sake, we focus on vulnerability to harm (Experiments 1 and 2) and moral standing (Experiment 3) in this paper and leave the additional items for future discussion.

In the “mental life” phase of the experiment, each participant evaluated their target’s ability to engage in four higher-order mental capacities (i.e., self-control, memory, communication, planning) and four experiential capacities (anger, fear, hunger, happiness), drawn from Gray et al., 2007. Like the transgression phase, we asked each of these questions initially within 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.

At the end of the experiment, we provided participants with a confirmation check (“Can you remind me, which of these characters was pushed to the ground and called a mean name?”).

Data preparation. To create a continuous scale for each variable, we coded “No” responses as 0, “a teeny bit” as 1, “a little bit” as 2, and “a lot” as 3.

Results & Discussion

To examine whether children’s evaluations of vulnerability to harm varied across targets, we submitted our data to a 3 (Condition: Robot, Human, Toy Bear) x 2 (Physical, Non-physical) mixed-model ANOVA with transgression type as a within-subjects factor. Children in our sample distinguished between these targets, $F(2, 225) = 10.01, p < .001$, rating the human boy as the most capable of suffering ($M = 2.78, SD = .66$), followed by the robot ($M = 2.40, SD = 1.00$), and the toy bear ($M = 2.04, SD = 1.27$). (Note, however, that the robot and toy bear targets did not differ significantly in harm vulnerability ratings, $t(134.65) = 1.94, p = .055$.) This provides some evidence that children, on the whole, consider robots as capable of at least *some* degree of suffering (i.e., 73 of the 116 children gave responses other than “0”).¹

We observed a similar pattern for mental life evaluations by target, evidenced by a separate ANOVA, $F(2, 915) = 97.73, p < .001$. Children ascribed more mental life to the human target ($M = 2.38, SD = 1.00$), as compared to the robot ($M = 1.97, SD = 1.23$) and the toy bear ($M = 1.07, SD = 1.33$). All targets differed significantly from one another ($ps < .001$). We also examined whether these targets differed by mental capacity type (i.e., agency vs. experience), in light of the proposed relationship between experiential capacities and moral standing (Gray et al., 2007). Here, we observed an interaction, $F(2, 912) = 10.38, p < .001$. Children distinguished

between the experiential capacities of the human boy, the robot, and the toy bear, $F(2, 455) = 40.46, p < .001$, but the human boy and robot targets were rated alike in terms of their higher-order (“agentic”) capacities (e.g., being able to remember things), $t(318) = -.61, p = .54$.

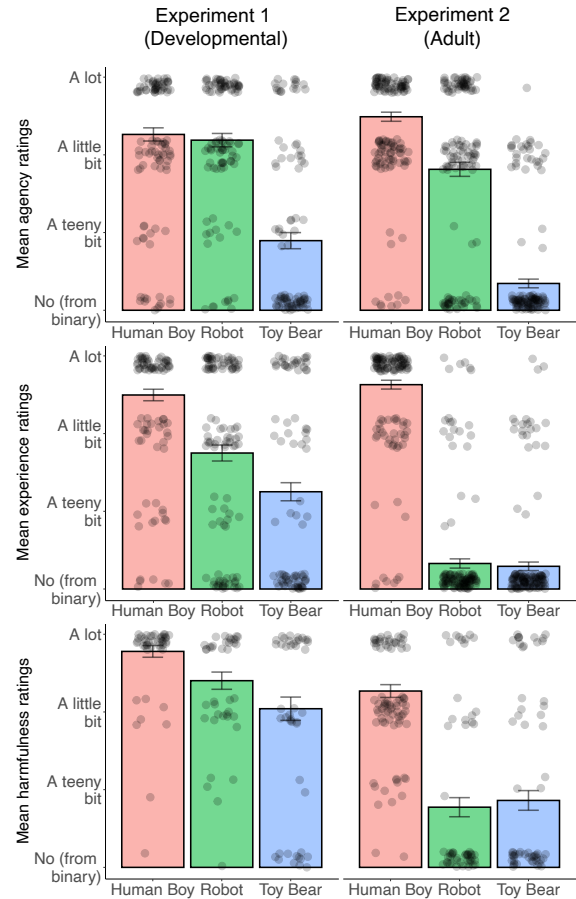


Figure 2: Mean ratings of agency (top panel), experience (middle panel), and vulnerability to harm (bottom panel) from Experiment 1 (developmental; left panel) and Experiment 2 (adult; right panel). Error bars represent +/- standard error of the mean.

Critically, children’s beliefs about mental life tracked with their judgments about vulnerability to harm: Experience ($\beta = .75, t = 8.03, p < .001$) and agency ($\beta = .43, t = 4.55, p < .001$) predicted children’s ratings about suffering. We see this as converging with existing accounts of moral standing (Schein & Gray, 2018). Our readiness to ascribe mental life to artificial minds seems critical for determining their vulnerability to harm.

Age effects. To explore whether these patterns shifted over the course of development, we analyzed the data in two ways. First, we categorized participants by two age groups (“younger” = 4 - 8.99 years of age, “older” = 9 - 13.99 years of age; 72 children categorized as “younger”, 44 as “older”) and analyzed the data using a mixed-effects model. Again,

¹The results for this experiment (as well as Experiments 2 and 3) did not meaningfully differ by whether a transgression was physical or non-physical in nature.

because we did not evenly recruit children by age (e.g., only one participant was 13 at the time of the experiment), there are some benefits to testing for potential age effects via age-based categories rather than as a continuous variable. Still, we did not observe an interaction between target assignment and age on children's ratings of harmfulness, $F(2, 109) = 1.16, p = .32$.

To determine whether (1) there were simply no age effects to be found, or (2) we were under-powered to detect potential age effects, we prepared Experiment 3 to further examine these potential relationships.

Experiment 2

In Experiment 1, we gained a preliminary understanding of how children relate harmfulness and mental life across a range of targets. In particular, children seemed to place the robot target in an "intermediary" space between the human boy and toy bear targets — both in regard to mental life and vulnerability to harm. Strikingly, these distinctions appear to be driven by children denying that robots were capable of experience and suffering (e.g., children rated humans and robots as equivalently agentic).

In Experiment 2, we provided the same experiment to a sample of adults. This sheds light on whether the patterns observed in Experiment 1 persist beyond childhood.

Method

Participants. We recruited 158 participants from Amazon Mechanical Turk (anticipating a sample of approximately 125, after exclusions). We excluded only a single participant for failing an attention check, leaving 157 in our final sample ($M_{age} = 33.8, SD_{age} = 9.77$; 79 identified as male, 71 identified as female, 1 identified as nonbinary). 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.

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

Results

We preregistered two predictions: First, consistently with the literature on speciesism, adults would rate the human target as having the most mental life and being the most vulnerable to harm, over and above the robot and toy bear targets. Second, adults would rate the robot as having more mental life than the toy bear (while also rating the robot and toy bear as similarly vulnerable to harm).

These predictions contrast with the developmental pattern observed in Experiment 1. Here, we anticipated that adults would not see robots as a moral "intermediary." Despite accepting that robots have some degree of mental life, adults would deny robots the capacity to suffer.

To test these hypotheses, we ran a series of mixed-model ANOVAs. Like children, adults distinguished between all tar-

gets in evaluating susceptibility to harm, $F(2, 301) = 56.53, p < .001$. As anticipated, the human boy was seen as the most capable of suffering ($M = 2.27, SD = .84$), with no differences between the robot ($M = .78, SD = 1.24$) and toy bear targets ($M = .86, SD = 1.27$), $t(200.76) = -.49, p = .62$. Further, although the human target, again, was rated highest in mental life (for both agency, $M = 2.49, SD = .85$, and experience, $M = 2.63, SD = .82$), adults also distinguished between the mental lives of the robot and toy bear. The adults in this sample rated the robot and bear as equivalently non-experiential, $t(401.49) = .45, p = .65$, and yet, on average, also rated the robot as having some degree of agency ($M = 1.81, SD = 1.29$).

This paints a clear developmental picture. Unlike the children in Experiment 1, adult participants denied that robots were capable of suffering — a pattern which was strongly predicted by the denial of experience-related mental capacities, $\beta = .75, t = 10.04, p < .001$, but not agency-related capacities, $\beta = -0.11, t = -1.38, p = .17$.

Experiment 3

In a final experiment, we gauged children's beliefs about the moral 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 see this as a stronger test of moral standing beliefs, as one might endorse that an entity *can* suffer without necessarily *caring* that the entity suffers.

We improved this experiment in two additional ways: First, in light of the differences between the developmental sample in Experiment 1 and the adult sample in Experiment 2, 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 that this provides important insight into how children compare these targets to one another when forming judgments.

Method

Participants. In light of a power analysis, we collected data from 161 children (ages 4 - 9.99). This allowed us to test for a medium-sized ($f = .25$) three-way interaction with 95% power. Before analysis, we removed data from 9 children (due to failing embedded manipulation checks). Our final sample consisted of 152 children ($M_{age} = 7.56, SD_{age} = 1.66$; 74 identified as male, 75 identified as female, 3 unidentified). Each child received a small prize for their participation.

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. All participants responded to all vignettes. We removed the vulnerability to harm item (e.g., "Do you think being pushed onto the ground hurt Drew?") and replaced it with an updated moral standing 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. Given that this experiment was entirely within-subjects and provided to children, we opted to include only 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).²

Results

We preregistered a set of three predictions. The first concerned moral standing. We anticipated that (1) as children aged, they would ascribe greater moral standing to the human boy than to the robot and toy bear targets. We predicted that younger children would discriminate between the three targets to a lesser degree. This is consistent with the speciesism endorsed by the adults in Experiment 2: Older children may find it “more okay” to transgress against the robot and toy bear targets.

The second and third predictions concerned mental life. We predicted that older children would ascribe heightened agency (2) and experience (3) to the human boy target (over the robot and toy bear), whereas younger children would discriminate between targets to a lesser degree. Together, these predictions align with our existing data: Adults clearly consider robots to be less experiential and less vulnerable to harm than do children. Here, we explore whether this presumed developmental effect is apparent in an additional measure of moral standing.

We fit a mixed-effects model (setting participant as a random effect) to examine whether children’s moral standing beliefs shifted over development, and whether this varied by target assignment. As predicted, children’s moral standing beliefs varied by age and target, $F(2, 18,546) = 130.1, p < .001$. This is to say that, in the bottom panel of Figure 3, the age trajectory for the human boy target (in red) differed from both the robot (in green, $\beta = -0.06, t = -9.51, p < .001$) and toy bear (in blue, $\beta = -.10, t = -16.04, p < .001$). This converges with findings from Experiments 1 and 2. Like adults, older children prioritized human moral standing, seeing it as “less okay” to transgress against humans (as compared to both non-human targets).

This is evident also in contrasts between targets. Though the youngest children in our sample (collapsing across ages 4 through 6) did distinguish between the moral standing of the human boy target and robot target, $t(43) = 2.24, p < .04$, the human boy target and the toy bear target, $t(43) = 4.85, p < .001$, and the robot target and toy bear target, $t(43) = 4.34, p < .001$, these effects were amplified amongst children between ages 7 and 9 (human-robot: $t(80) = 3.69, p < .001$,

²We 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 the participants’ choices over the computer. These shifts in phrasing did not impact the results of the experiment.

human-toy: $t(80) = 8.22, p < .001$, robot-toy: $t(80) = 7.10, p < .001$).

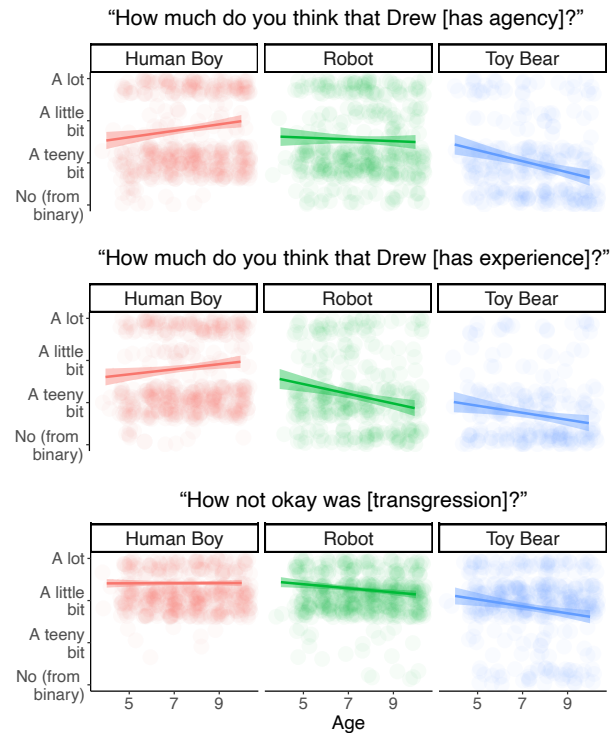


Figure 3: Scatterplot of children’s ratings of agency (top panel), experience (middle panel), and moral standing (bottom panel) from Experiment 3 (with 95% confidence intervals). The x-axis denotes participant age.

Age effects also emerged for mental life. With age, the children in our sample ascribed less experience (robot, $\beta = -.19, t = -28.70, p < .001$; toy bear, $\beta = -.15, t = -21.68, p < .001$) and agency (robot, $\beta = -.11, t = -17.24, p < .001$; toy bear, $\beta = -.24, t = -35.82, p < .001$) to the robot target and to the toy bear, as compared to the human boy. This suggests that children demoralize both the robot and toy bear over the course of childhood. Both of these metrics correlated with moral standing evaluations (agency: $r = .48, 95\% \text{ CI } [.40, .55], p < .001$; experience, $r = .52, 95\% \text{ CI } [.45, .59], p < .001$). This converges with a wide range of papers connecting mental life with moral standing: Younger children’s tendency to anthropomorphize may correspond with maintaining a more expansive moral circle (Wilks et al., 2021).

General Discussion

Despite robots’ increasing presence in the human social world, little work has addressed people’s beliefs about the moral standing of artificial intelligence. In three experiments, we examined children and adults’ beliefs about the moral standing of artificial life. We find that, on the whole, children place robots in an “intermediary” category between natural kinds and artifacts (e.g., Kahn, Gary, & Shen, 2013). Chil-

dren seem to believe that robots can suffer (Experiment 1), and that it is morally wrong to transgress against robots (Experiment 3) — though these beliefs diminish with age. By contrast, adults, deny that robots have the ability to suffer (Experiment 2).

All of this coincides with beliefs about mental life: Early in development, children endorse that robots have rich mental lives, capable of both agency and experience. Older children and adults, by contrast, tend to deny that artificial minds are capable of experience (Experiments 2 and 3). Tentatively, this suggests that “speciesism,” granting privileged moral standing to humankind, may be learned and exclusive to later childhood and adulthood.

Limitations

There are a number of limitations to this set of experiments. First, we focused on a narrow set of stimuli and collected data exclusively from the United States. We recommend testing these intuitions with broader stimuli and across other populations (e.g., Takahashi, Ban, & Asada, 2016) before interpreting these results as applying universally. Given the potential role of social learning in the formation of beliefs about moral standing (e.g. Wilks et al., 2021), it is possible that developmental trends could emerge differently across different cultures. These potential differences may constrain the generalizability of our results.

A second concern has to do with children’s relatively high ratings of mental life, vulnerability to harm, and moral standing for the toy bear. This was unexpected; both experimental evidence (e.g., Carey, 1985) and everyday experience suggest that children do not actually think that toys are sentient moral entities. When responding to our vignettes, children may have instead been pretending or play-acting, the way that one does when reading stories such as the “Berenstain Bears.” This raises the concern that children were engaging in pretense when evaluating the robot target as well. On the other hand, children gave a similar pattern of responses for the human boy. Here, they plainly were not pretending: It seems obvious that children truly believe that other children have mental states and moral standing. Perhaps they think the same of robots. In future work, we plan to explore in more detail the nature of children’s responses to robot stimuli.

Relatedly, we recognize that both children and adults may have interpreted the phrasing of our harm vulnerability and moral standing items in an unintended manner (e.g., endorsing that it was “not okay” to transgress against a robot because the action causes property damage, rather than an offense to moral standing). Given that people likely interpret these items in the intended manner when directed towards the human target (e.g., believing that it is “not okay” to transgress against a human because the action is an actual offense to moral standing), this potential difference of interpretation in light of target further suggests a distinction between human and robot moral standing. Robots may be perceived to have exclusively extrinsic moral standing (Zimmerman & Bradley, 2019). It would be valuable for future research to disentangle

these possible mechanisms.

Future Directions

We see the present set of experiments as an initial 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 is unclear. Do they believe that robots have moral rights? What happens 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 these possibilities. Second, we focused our scope to people’s evaluations of robots as *victims* of harm. It remains unknown whether these patterns extend to evaluations of robots as *perpetrators*. To speculate, we think it would be interesting if robots maintained an “intermediary” rank for moral responsibility — being evaluated as less accountable than a human for a moral transgression (but more accountable than a non-agent; Kahn et al., 2012).

This work also raises interesting questions concerning mechanism. Here, we highlight the relationship between mental life and moral standing, echoing a large existing literature in moral psychology (Gray et al., 2007; Schein & Gray, 2018). There are a host of other possible moderators which may contribute to diminished perceptions of moral standing. One such possibility is outgroup degradation. Given that speciesism tracks with a range of human-based social prejudices (Everett, Caviola, Savulescu, & Faber, 2019), it may be that adults’ apprehension toward granting non-humans with moral privileges stems in part from distaste for “the other” (e.g., artificial intelligence as part of a non-human outgroup). Future work would do well to investigate these possibilities.

References

- Bloom, P., & Harris, S. (2018). It’s Westworld. What’s Wrong With Cruelty to Robots? *The New York Times*.
- Brink, K. A., Wellman, H. M., & Gray, K. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development, 90*, 1202–1214.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Caviola, L., Everett, J. A. C., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology, 116*(6), 1011–1029. doi: 10.1037/pspp0000182
- Everett, J. A. C., Caviola, L., Savulescu, J., & Faber, N. S. (2019). Speciesism, generalized prejudice, and perceptions of prejudiced others. *Group Processes & Intergroup Relations, 22*(6), 785–803. doi: 10.1177/1368430218816962
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619. doi: 10.1126/science.1134475
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cogni-*

- tion, 125(1), 125–130. doi: 10.1016/j.cognition.2012.06.007
- Jipson, J. L., & Gelman, S. A. (2007). Robots and Robots: Children's Inferences About Living and Nonliving Kinds. *Child Development*, 78(6), 1675–1688. doi: 10.1111/j.1467-8624.2007.01095.x
- Jordan, 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 the Sciences*, 111, 12710–12715.
- Kahn, P. H., Gary, H. E., & Shen, S. (2013). Children's Social Relationships With Current and Near-Future Robots. *Child Development Perspectives*, 7(1), 32–37. doi: <https://doi.org/10.1111/cdep.12011>
- 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). doi: 10.1037/a0027033
- Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., ... Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. *Proceedings of the 6th international conference on Human-robot interaction*, 159–160. doi: 10.1145/1957656.1957710
- Melson, G. F., Kahn, P. H., Beck, A., Friedman, B., Roberts, T., Garrett, E., & Gill, B. T. (2009). Children's behavior toward and understanding of robotic and living dogs. *Journal of Applied Developmental Psychology*, 30(2), 92–102. doi: 10.1016/j.appdev.2008.10.011
- Nigam, M. K., & Klahr, D. (2000). If robots make choices, are they alive?: Children's judgments of the animacy of intelligent artifacts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 22(22).
- Nijssen, S. R. R., N., M. B. C., van Baaren, B., R., & Paulus, M. (2019). Saving the robot or the human? Robots who feel deserve moral care. *Social Cognition*, 37, 41–56.
- Risse, M. (2019). Human rights and artificial intelligence: An urgently needed agenda. *Human Rights Quarterly*, 41.
- Schein, C., & Gray, K. (2018). The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm. *Personality and Social Psychology Review*, 22, 108886831769828. doi: 10.1177/1088868317698288
- Singer, P. (2009). Speciesism and Moral Status. *Metaphilosophy*, 40(3-4), 567–581. doi: <https://doi.org/10.1111/j.1467-9973.2009.01608.x>
- 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, 104656. doi: 10.1016/j.jecp.2019.06.009
- Takahashi, H., Ban, M., & Asada, M. (2016). Semantic Differential Scale Method Can Reveal Multi-Dimensional Aspects of Mind Perception. *Frontiers in Psychology*, 7. doi: 10.3389/fpsyg.2016.01717
- Weisberg, D. S. (2015). Pretend play. *WIREs Cognitive Science*, 6(3), 249–261. doi: <https://doi.org/10.1002/wcs.1341>
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Children's intuitions about the structure of mental life. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.
- Wilks, M., Caviola, L., Kahane, G., & Bloom, P. (2021). Children Prioritize Humans Over Animals Less Than Adults Do. *Psychological Science*, 32(1), 27–38.
- Zimmerman, M. J., & Bradley, B. (2019). Intrinsic vs. Extrinsic Value. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University.