Community Futures With Morally Capable Robotic Technology

Terran Mott
MIRRORLab
Colorado School of Mines
Golden, CO, USA
terranmott@mines.edu

Tom Williams
MIRRORLab
Colorado School of Mines
Golden, CO, USA
twilliams@mines.edu

ABSTRACT

In this paper, we consider the technology literacy needs of human communities pertaining to robots' moral agency and moral competency. We consider how user communities will need to make judgements about when to attribute moral agency to robots, create policies based on this understanding, and make choices on behalf of others about robots' involvement in their lives. We propose that the technology literacy benchmarks in Project 2061 offer a compelling set of guidelines for empowering users to make informed, competent judgments about how morally capable social robots ought to participate in their lives.

KEYWORDS

robot ethics, technology literacy

1 INTRODUCTION

Social will inevitably encounter a variety of morally fraught situations. Robots may be given unethical commands [21]. They may be bystanders to abusive language or harmful behavior [23, 24]. They may even have the opportunity to confront societal biases [46, 47]. Researchers have argued that robots need to be able to appropriately respond to these situations, as failure to do so will, at minimum, risk tacit acceptance of these observed norm violations [7, 23].

In order to respond to these violations, robots may need to have capabilities associated with *agency*, such as autonomy, interactivity, and adaptability [12]. By having the language capabilities needed to respond to these violations, robots may additionally be seen as capable of morally and socially consequential actions, giving them moral and social agency [22] as well. Critically, moral and social agency may come with expectations of moral competence [29].

Researchers have explored robotic moral agency and competence from a variety of perspectives—from philosophical and ethical theories [12, 49, 50], moral psychology [25, 29] and algorithm design [3, 44]. Human-robot interactions involving moral norms are impacted by roles and relationships [43], gender [23, 47], affect [6], and linguistic politeness [15, 16, 37]. This body of work has established many considerations for how to maximize the benefits and minimize the harms in morally fraught robot interactions. However, it is also critical for roboticists to consider the implications of morally capable social robots beyond the level of individual interactions.

Technology and society continually and mutually shape [36] and mediate [40] one another. Researchers [11, 34] and policymakers [14, 45] have thus called for more broader sociotechnical perspectives on our future with robots. Such perspectives emphasize social, legal, emotional, or institutional externalities. In line with this need for broader sociotechnical perspectives, we ask: how might we support communities in making good judgements about the moral agency and moral competence of robots?

Fundamentally, humans must make judgements about the extent to which robots are social, moral, and intelligent others [42]. One way to support users in making accurate judgements about a robot's capabilities is to follow the design principle of *transparency*, the idea that technology can communicate its inner workings, capabilities, and limitations to users [2]. HRI researchers [1, 41], as well as policymakers [10], have explored how transparent systems can support users. This kind of information can help people accept robots [26], maintain better Situation Awareness [5, 9], build accurate mental models [4, 27, 48], and calibrate their trust [1, 35].

Those who are more informed of robots' capabilities and limitations can make better judgements about how to understand, use, and trust them. However, for many future robot users, decisions about whether and when to use and trust robots may happen before those users have the chance to interact with a given robotic technology. People will need to choose which robotic platforms to purchase, and what kind of initial role that platform will be given within their use context. These judgements must be made before users have much experience interacting with the system and observing it at the user's level of abstraction [22]. Even if a robot is transparent about its capabilities during interactions, users may not have access to this information to help them make initial decisions about purchasing, using, and trusting the system.

We argue that supporting user communities and institutions to evaluate morally capable social robots can be understood through the lens of technology literacy. We explore what it might mean for a user community to be technologically literate with respect to artificial moral agents and to make good judgements and decisions about the role morally agentic robots can or should play in their spaces. We ask the question: *How can we support the users of morally capable robots in building accurate understanding, calibrating trust, and making decisions at the level of institutions, communities and societies?* We then present a set of selected technology literacy benchmarks, adapted from those in the American Association for the Advancement of Science "Project 2061" [31]. A community in which members are technologically literate with respect to artificial moral agents can make good decisions about whether, when, and how such technology should participate in their spaces.

2 USER COMMUNITIES MAKE JUDGEMENTS ABOUT MORAL AGENTS

In the future, people ought to be empowered to make good decisions about whether, when, and how to engage with morally capable social robots. Technology literacy can prepare user institutions and communities to make good judgements about such technology, before they have the opportunity to interact consistently with it. For example, consider the community of teachers, administrators, and parents at a school. Suppose this group is responsible for making decisions about which robotic platform to purchase for their classrooms (or whether to purchase one at all). They must weigh the benefits of such technology to provide assistance and support to children [18, 38] against the potential harms, such as deception [19, 39]. Suppose the school is particularly sensitive to the potential for moral harm, and prioritizes selecting a classroom robot with moral competence. They are aware that their classroom robots will likely confront moral norm violations, and wish to select a robot that can model fair and morally appropriate behavior.

An informed, technologically literate perspective can help those in the school make reasonable attributions of robots' moral agency, understand their implications, and create effective policies about robots. Here, we explore three fundamental decisions within this process that stakeholders must make. First, they must decide whether to attribute moral agency to a robot and form expectations of its moral competence. Then, they must decide what role the robot should play in their particular context.

2.0.1 Evaluating News & Advertising. Stakeholders at the school will need to make initial judgements about which robot is right for them. Importantly, these judgments and purchasing decisions will happen before they have the benefit of in-person experience interacting with and observing the system at the user's level of abstraction [22]. This means that stakeholders will likely rely on news and advertising about robotic technology in order to form their initial mental models of its capabilities—deciding which available platforms seem to have moral agency, and trying to decide which one has the moral competence to be successful navigating the moral norms and fraught situations it might encounter in their classrooms. Technology literacy can help those at the school assess the accuracy and credibility of these sources. This may involving identifying reasonable or sensationalized news, or identifying valid or suspect claims in advertising for new robotic products.

For example, consider how those with a good understanding of the strengths and weaknesses of social technology can identify suspicious claims in advertising. Suppose a school administrator is compiling their initial list of options for potential classroom robots. Some of the claims made by such companies might be exaggerated. They might claim their robots always tell when humans are being polite or rude. They might claim their robot can read children's facial expression to tell what they are feeling. Both of these capabilities might be important input to help a robot's moral reasoning; however, both politeness detection [17, 20] and children's emotion classification [8] are incredibly difficult, unreliable problems. Strong technology literacy, such as the intuition that robots often struggle with the more nuanced and context-dependent parts of social interaction, may help the administrator avoid taking these claims at face value, and encourage their team to dig deeper into the robots' real capabilities. Technology literacy surrounding moral agency and moral competence could similarly allow communities to understand whether a robot could *really* respond to violations; and to understand the effects of placing such a robot into their classroom, regardless of whether or not it actually intervenes.

2.0.2 Creating & Revising Policies. Stakeholders at the school will likely also set up procedures and policies for how to use their morally capable social robots before interacting with them. Their understanding of artificial agents and judgements about moral agency will inform these policy decisions about blame and accountability [28, 33]. On the broadest scale, they may have the opportunity to vote on laws or policies concerning the regulation of robotic technology. They may also be involved in creating, revising, or approving policies at an institutional scale.

For example, once the school has selected a robotic platform to purchase as their in-classroom robotic companions, they will likely need to create some basic policies and procedures about whether and how to use the robots. Technology literacy can help them create effective polices. For example, teachers may understand that they should temper their trust in robots when it comes to navigating certain moral norms. There may be some morally fraught situations that teachers trust robots to handle on their own, such as encouraging students to follow a conflict resolution procedure when they become upset during a collaborative task [24, 38]. However, there may also be other morally fraught situations that teachers do not trust the robot alone with, and would intervene or supervise. This may include much more complex situations with more potential for harm, such as instances of bigotry or bias [23, 46]. Technology literacy can help stakeholders create policies about what a morally capable social robot's role should be and when it should be trusted.

2.0.3 Making Choices on Behalf of Others. In the future, people may need to make decisions on behalf of other members of an institution, community, or society regarding the role of morally capable robotic technology. Families may need to make decisions about whether to invest in robotic assistance for a relative. Leaders of professional teams may be involved in decisions about social robots on behalf of their employees. Their understanding of robots' moral capabilities, and the limits of these capabilities, will guide such decisions. This may be especially true for vulnerable populations, who may not be able to make decisions for themselves.

For example, parents of children at the school might be given the opportunity to decide on behalf of their child if they consent to having their child interact with a classroom robot unsupervised. Technology literacy can help them accurately weigh the pros and cons of child-robot interaction, even if parents have little opportunity to interact with the robot themselves.

2.1 Project 2061 Technology Literacy Guidelines

In the previous section, we explored examples of the decisions that user institutions and communities may make about morally capable robotic technology, before having the chance to build mental models of its capabilities through real interaction. If roboticists are to support communities in making good judgements, then they should support these communities technology literacy. Here, we present a set of scientific literacy guidelines from Project 2061, and propose that they offer a strong foundation upon which to imagine technologically literate user communities who are prepared to make good judgments about morally capable social robots.

Project 2061 is a scientific literacy initiative by the American Association for the Advancement of Science [30, 32]. The project began after the 1985 passage of Halley's Comet, and aimed to prepare children to evaluate the scientific and technological changes they should expect to see before the comet returns in 2061 [31]. Project

2061 has since expanded to include several technology education initiatives, and the guidelines themselves were updated in 1993 and 2009 to reflect technological advancement [13].

Here, we present a subset of Project 2061 guidelines on "The Nature of Technology." We selected these guidelines for relevance (For example, we discarded those relating to the job opportunities available in technology). The complete list is available in [13]. We argue that people who have a strong understanding of these concepts would be prepared to make competent judgements about the role of morally capable social robots in their community.

(1) In designing a device or process, thought should be given to how it will be manufactured, operated, maintained, replaced, and disposed of and who will sell, operate, and take care of it. The costs associated with these functions may introduce yet more constraints on the design.
(2) The value of any given technology may be different for different groups of people and at different points in time.
(3) Complex systems have layers of controls. Some controls operate particular parts of the system and some control other controls. Even fully automatic systems require human control at some point.
(4) Risk analysis is used to minimize the likelihood of unwanted side effects of a new technology. The public perception of risk may depend, however, on psychological factors as well as scientific ones.
(5) The more parts and connections a system has, the more ways it can go wrong. Complex systems usually have components to detect, back up, bypass, or compensate for minor failures.
(6) Social and economic forces strongly influence which technologies will be developed and used. Which will prevail is affected by many factors, such as personal values, consumer acceptance, patent laws, the availability of risk capital, the federal budget, local and national regulations, media attention, economic competition, and tax incentives.
(7) In deciding on proposals to introduce new technologies or curtail existing ones, some key questions arise concerning possible alternatives, who benefits and who suffers, financial and social costs, possible risks, resources used (human, material, or energy), and waste disposal

3 DISCUSSION AND FUTURE IMPLICATIONS

The Project 2061 technology literacy guidelines are advantageous in an HRI setting because they align strongly with other work exploring the sociotechnical implications of robots. For example, the selected guidelines encourage the understanding that economic or social forces influence technology, and vice versa [36]. They also emphasize how technology may not evenly distribute its benefits and burdens among stakeholders [34]. They highlight how there is always human involvement in designing technology, even when that technology is highly autonomous or complex.

The Project 2061 guidelines may help a user community to effectively evaluate the moral capabilities of robots. In particular, they may guide non-technologists in thinking critically about robots' capabilities in an economic and social context, before making the financial commitment to purchase them. They may also encourage stakeholders to emphasize the role and responsibility of technologists, and to hold them accountable for the potential failures of autonomous systems. However, these guidelines are broad. They lead to new research questions about how HRI researchers may more specifically explore technology literacy in the context of artificial moral agents. Inspired by Project 2061, roboticists can explore future research questions about supporting user communities of morally capable social robots, such as:

- Can we create technology literacy guidelines specifically about moral agency, robotic moral reasoning, and the ability of robots to take moral actions?
- What policies might effectively regulate the way future companies advertise the moral capabilities of their robots?
- How can technologists support communities in making good judgements about whether or when to use morally agentic robotic technology?

4 CONCLUSION

We consider the technology literacy needs of communities pertaining to robots' moral agency and moral competency. We considered how user communities will need to make judgements about when to attribute moral agency to robots, create policies based on this understanding, and make choices on behalf of others about robots' involvement in their lives. We propose that the technology literacy guidelines in Project 2061 offer a compelling set of guidelines for empowering users to informed, competent judgments about how morally capable social robots ought to participate in their lives.

ACKNOWLEDGMENTS

This work was funded in part by NSF CAREER award IIS-2044865, Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research, and award FA9550-23-1-0036 from the United States Air Force Office of Scientific Research.

REFERENCES

[1] Victoria Alonso and Paloma de la Puente. 2018. System Transparency in Shared Autonomy: A Mini Review. *Frontiers in neurorobotics* (2018). doi.org/10.3389/fnbot.2018.00083

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1078–1088.

[3] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. 2021. An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1911–1918. https://doi.org/10.1109/IROS51168.2021.9636434

[4] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L. Glassman. 2022. Revisiting Human-Robot Teaching and Learning Through the Lens of Human Concept Learning. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 147–156.

[5] Michael W. Boyce, Jessie Y.C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. 2015. Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (Portland, Oregon, USA) *(HRI'15 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 179–180. https://doi.org/10.1145/2701973.2702059

[6] Gordon Briggs and Matthias Scheutz. 2014. How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress. *International Journal of Social Robotics* 6 (2014), 343–355.

[7] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. 2022. Why and how robots should say 'no'. *International Journal of Social Robotics* 14, 2 (2022), 323–339.

[8] De'Aira Bryant and Ayanna Howard. 2019. A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 377–382. https://doi.org/10.1145/3306618.3314284

[9] Jessie Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. Situation Awareness–Based Agent Transparency. *US Army Research Laboratory* (01 2014).

[10] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI.* Publications Office. https://doi.org/doi/10.2759/346720

[11] Katie Darling. 2021. *The New Breed: What Our History with Animals Reveals About Our Future with Robots.* Henry Holt and Company.

[12] Luciano Floridi and J.W. Sanders. 2004. On the Morality of Artificial Agents. *Minds and Machines* 14 (08 2004), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[13] American Association for the Advancement of Science. 2009. The Nature of Technology. (2009). http://www.project2061.org/publications/bsl/online/index.php?chapter=3#B4

[14] Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, and Jonathan M. Smith. 2017. Toward Algorithmic Transparency and Accountability. *Commun. ACM* 60, 9 (aug 2017), 5. https://doi.org/10.1145/3125780

[15] Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents. *Cognitive Science* (2017).

[16] Victoria Groom, Jimmy Chen, Theresa Johnson, F. Arda Kara, and Clifford Nass. 2010. Critic, Compatriot, or Chump? Responses to Robot Blame Attribution. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, 211–218.

[17] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 52 (dec 2017), 14 pages. https://doi.org/10.1145/3134687

[18] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting. *ACM/IEEE International Conference on Human-Robot Interaction* 2015 (03 2015), 83–90. https://doi.org/10.1145/2696454.2696479

[19] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *The American psychologist* 74, 2 (July 2019). https://doi.org/10.1037/amp0000307

[20] Nasif Imtiaz, Justin Middleton, Peter Girouard, and Emerson Murphy-Hill. 2018. Sentiment and Politeness Analysis Tools on Developer Discussions Are Unreliable, but so Are People. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering* (Gothenburg, Sweden) *(SEmotion '18)*. Association for Computing Machinery, New York, NY, USA, 55–61. https://doi.org/10.1145/3194932.3194938

[21] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 499–505. https://doi.org/10.1145/3306618.3314241

[22] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (2021). https://doi.org/10.3389/frobt.2021.687726

[23] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 559–567. https://doi.org/10.1145/3319502.3374831

[24] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 229–236. https://doi.org/10.1145/2696454.2696460

[25] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, USA, 10–18. https://doi.org/10.1145/3434074.3446908

[26] Johannes Kraus, Franziska Babel, Philipp Hock, Katrin Hauber, and Martin Baumann. 2022. The trustworthy and acceptable HRI checklist (TA-HRI): questions and design recommendations to support a trust-worthy and acceptable design of human-robot interaction. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)* (08 2022), 1–21. https://doi.org/10.1007/s11612-022-00643-8

[27] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 463–464. https://doi.org/10.1109/HRI.2016.7451807

[28] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 715, 11 pages. https://doi.org/10.1145/3411764.3445284

[29] Bertram F Malle and Matthias Scheutz. 2020. Moral competence in social robots. In *Machine ethics and robot ethics.* Routledge, 225–230.

[30] William F. McComas. 2014. *The Atlas of Science Literacy.* SensePublishers, Rotterdam, 9–9. https://doi.org/10.1007/978-94-6209-497-0_8

[31] William F. McComas. 2014. *Benchmarks for Science Literacy.* SensePublishers, Rotterdam, 12–12. https://doi.org/10.1007/978-94-6209-497-0_11

[32] J. Meinwald, J.G. Hildebrand, American Academy of Arts, and Sciences. 2010. *Science and the Educated American: A Core Component of Liberal Education.* American Academy of Arts and Sciences. https://books.google.com/books?id=WVP6NAEACAAJ

[33] Sari R. R. Nijssen, Barbara C. N. Müller, Rick van Baaren, and Markus Paulus. 2019. Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Social Cognition* (2019).

[34] Anastasia K. Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* 969–976. https://doi.org/10.1109/RO-MAN53752.2022.9900805

[35] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in Human–Agent Systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (nov 2019), 673–705. https://doi.org/10.1007/s10458-019-09408-y

[36] S. Sabanovic. 2010. Robots in Society, Society in Robots. *International Journal of Social Robotics* 2 (12 2010), 439–450. https://doi.org/10.1007/s12369-010-0066-7

[37] Maha Salem, Micheline Ziadee, and Majd Sakr. 2014. Marhaba, How May i Help You? Effects of Politeness and Culture on Robot Acceptance and Anthropomorphization. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 74–81. https://doi.org/10.1145/2559636.2559683

[38] Solace Shen, Petr Slovak, and Malte F. Jung. 2018. Stop. I See a Conflict Happening.: A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 69–77. https://doi.org/10.1145/3171221.3171248

[39] Caroline L. van Straten, Jochen Peter, Rinaldo Kahne, and Alex Barco. 2021. The wizard and I: How transparent teleoperation and self-description (do not) affect children's robot perceptions and child-robot relationship formation. *AI & SOCIETY* (April 2021). https://doi.org/10.1007/s00146-021-01202-3

[40] Peter-Paul Verbeek. 2015. Toward a theory of technological mediation. *Technoscience and postphenomenology: The Manhattan papers* 189 (2015).

[41] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* 10, 3, Article 27 (jul 2021), 24 pages. https://doi.org/10.1145/3457188

[42] Kara Weisman. 2022. Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[43] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-Based Relational Norms. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 353–362.

[44] Ruchen Wen and Tom Williams. 2022. Hidden Complexities in the Computational Modeling of Proportionality for Robotic Norm Violation Response. In *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI).*

[45] Mark West, Rebecca Kraut, and Chew Han Ei. 2019. I'd blush if I could: closing gender divides in digital skills through education. (2019).

[46] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Bršcić, Iolanda Leite, and Tom Williams. 2022. Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 120–129.

[47] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 29–37. https://doi.org/10.1145/3434074.3446910

[48] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. In *Proceedings of the IJCAI Workshop on Ethics for Artificial Intelligence*. IJCAI 2016 Ethics for AI Workshop ; Conference date: 09-07-2016 Through 09-07-2016.

[49] Qin Zhu, Tom Williams, and Ryan Jackson. 2018. Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective. In *Brain Based and Artificial Intelligence*.

[50] Qin Zhu, Tom Williams, and Ruchen Wen. 2021. Role-based morality, ethical pluralism, and morally capable robots. *Journal of Contemporary Eastern Asia* 20, 1 (2021), 134–150.