Fairness and Transparency in Human-Robot Interaction

Houston Claure
Mechanical Engineering
Cornell University
Ithaca, New York
hbc35@cornell.edu

Mai Lee Chang
Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas
mlchang@utexas.edu

Seyun Kim
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh,Pennsylvania
seyunkim@andrew.cmu.edu

Daniel Omeiza
Department of Computer Science
University of Oxford
Oxford, England
daniel.omeiza@cs.ox.ac.u

Martim Brandão
Department of Informatics
King's College London
London, England
martim.brandao@kcl.ac.uk

Min Kyung Lee
School of Information
University of Texas at Austin
Austin, Texas
minkyung.lee@austin.utexas.edu

Malte Jung
Information Science
Cornell University
Ithaca, New York
mfj28@cornell.edu

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) | 978-1-6654-0731-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/HRI53351.2022.9889421

***Abstract*—As robots become more ubiquitous across human spaces, it is becoming increasingly relevant for researchers to ask the question, "how can we ensure that we are designing robots to be sufficiently equipped to treat people fairly?". This workshop brings together researchers across the fields of Human-Robot Interaction (HRI), fairness in machine learning, design, and transparency in AI to shed light on the relevant methodological challenges surrounding issues of fairness and transparency in HRI. In our workshop, we will attempt to identify synergies between these various fields. In particular, we will focus on how HRI can leverage these existing rich body of work to guide the formalization of fairness metrics and methodologies. Another goal of the workshop is to foster a community of interdisciplinary researchers to encourage collaboration. The complexity in defining fairness lies in its context sensitive nature, as such we look to the influx of definitions from the field of fairness in artificial intelligence, design, and organizational psychology to derive a set of definitions that could serve as guidelines for researchers in HRI.**

***Index Terms*—Fairness in HRI; Ethics in HRI; Transparency in AI**

I. Introduction

Fairness has been at the forefront of many recent discussions revolving the introduction of intelligent systems into decision making contexts. Some key concerns in this space involve removing underlying biases across the different stages of the machine learning pipeline that can enable negative consequences towards a protected group or individual [1]–[6]. Along a similar vein, researchers within the field of human-robot interaction (HRI) have recently begun exploring how robotic behavior can elicit different fairness considerations depending on the context in which the robot is deployed. This has given rise to a host of research questions revolving around themes of fairness and teamwork [7]–[9], navigation [10], and design [11] to name a few. Across this research, a broad range of definitions and metrics emerged highlighting the necessity for a deeper conversation about methods and measurements of fairness and transparency within HRI.

The recent push to explore fairness and intelligent systems has driven researchers to draw inspiration from traditional streams of fairness literature (e.g. organizational psychology [12], economics [13], sociology [14]). Specifically, there has been a host of works investigating how AI systems can produce statistical fairness [15] along with exploring how these algorithm's decisions are perceived [16], [17]. The findings of these works highlight the need for computational systems to understand the values and dynamics that exist between humans in the environment for their successful adoption [18]. As robots are being deployed in new frontiers, it is essential to continue investigating the relevant factors that influence fairness and transparency in a systematic way. Towards this goal, this workshop aims to investigate how fairness and transparency can be defined across different contexts and will explore the potential impact on shaping human relationships.

Through this workshop we will bring findings and understandings from a broad range of fields in an effort to shape an agenda for future directions in fairness and transparency within HRI. We will bring speakers across the quantitative and qualitative spaces in order to ensure a holistic discussion about: (i) current and existing works within the space, (ii) key methodological challenges, (iii) various relevant metrics and definitions, and (iv) best practices and techniques to explore fairness and transparency.

II. Fairness and Transparency in HRI

Within HRI, fairness has primarily been explored through design and decision making algorithms. Early findings highlight the social implications that the interpretation of fairness has on individuals across a variety of contexts such as multi human teams [7], [8], [19], [20] and navigation [10], [21]. The context dependent nature of fairness provides a challenge for researchers on established methods to study fairness as well as a set of metrics to apply. This challenge has made it essential to develop newer methods of evaluation. Some examples

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 30,2023 at 17:38:07 UTC from IEEE Xplore. Restrictions apply.

include works by Chang et al. [8] and Claure et al. [7] who have used games and video stimulations that mirror common contexts where robots are placed in resource allocation roles [19], [22]. Their work demonstrated how a robot's allocation decisions can influence team behavior and shape perceptions of trust towards the system. Fairness has also been explored through the lens of robot design and robot behavior [11]. Researchers have argued that how a robot is portrayed and designed can elicit fairness interpretations [23]. Taken together these works point towards important gaps in literature that need to be addressed in order to push the agenda on fairness and transparency in HRI. Such gaps include more in-the-wild experiments, exploring the effects of robot embodiment on fairness perceptions, and algorithms that enable a robot to learn human fairness.

Previous workshops have explored topics around fairness and transparency where they identified the need for better methods of evaluation for these concepts in HRI [24], [25]. This workshop will extend the findings from such workshops and focuses on discussions about methodological challenges and solutions that would benefit the broader HRI community. We will implement interdisciplinary approaches in order to draw expertise from different researchers both in academia and industry.

III. Workshop Overview

We propose a half-day workshop aiming to discuss the different practices and metrics that are relevant for researchers in HRI. By involving discussions from researchers across different spaces, we aim to create tools and definitions to advance the application of fairness and transparency. Upon the completion of the workshop, we will upload position papers to the website and continue a blog to ensure that the website acts as a repository for any new information revolving fairness in HRI. The blog will further push the ideas from the workshop and will store new metrics and definitions that have been used in the space of robotics. We additionally plan to create a working group to further foster a community of researchers across a broad spectrum of robotics to share ideas and encourage collaboration.

IV. Schedule and Activities

The half-day workshop will include experts from both industry and academia who will discuss current trends within the space of fairness and transparency. Specifically, we aim to bring in speakers such as Dr. Ayanna Howard, Dr. Cynthia Dwork, Dr. Solon Barocas, or Dr. Kate Tsui who can further speak towards best practices and definitions for fairness and transparency within HRI. We will include two separate break out sessions where participants will be split into groups to complete discussion and brainstorming sessions that will be moderated by the organizing committee. Finally, participants will be invited to present their selected accepted papers to explore different perspectives and facilitate discussions amongst participants. See a tentative schedule in Table I.

TABLE I
Proposed Schedule

Schedule	Topic
12:00 - 12:10	Introduction
12:10 - 12:40	Invited Speaker 1
12:40 - 13:10	Break Out Session 1: Discussion of Key Definitions and Metrics
13:10 - 13:20	Break
13:20 - 14:00	Paper Presentations
14:00 - 14:30	Invited Speaker 2
14:30 - 15:00	Break Out Session 2: Future Directions

A. List of Topics

Relevant topics of interest for this workshop include but are not limited to:

- Trustworthy AI
- Trust and Human-Robot Interaction
- Ethics implications in HRI
- Ethical design of robotic systems
- Age/race/gender-biased robots
- Transparency in HRI
- Human biases in HRI
- Development and study of fair machine learning models in robotics
- Interaction design and explainable AI
- Metrics for studying fairness
- Fairness in resource allocation
- Fairness in Human-robot teams

V. Audience and participation

Participants from the fields of HCI, HRI, psychology, and fairness in machine learning will be welcome to submit a 2-3 page position paper. We particularly will encourage individuals who are exploring the topics of fairness (both quantitative and qualitative) and transparency in robotics. These papers will be peer reviewed by committee members. We will request that at least one author must be present at the workshop in order to present during the workshop. Finally, we will recruit 20-25 participants via relevant mailing lists and social media.

VI. Organizers

The organizing team consist of researchers who focus on both the quantitative and qualitative aspects of fairness and transparency in AI and HRI.

Houston Claure is a Ph.D. Candidate at Cornell University in the Robots in Groups Lab. His research involves exploring the use human notions of fairness to shape robotic decisions within multi human teams towards the goal of optimizing team performance and cohesion.

Mai Lee Chang is a Ph.D. Candidate at the University of Texas at Austin in the Socially Intelligent Machines Lab. Her research goals are to enable robotic teammates to reason about task performance and fairness to achieve long-lasting human-robot partnerships.

Daniel Omeiza is a 3rd-year Ph.D. student at the University of Oxford, working on explainability in autonomous driving. He

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 30,2023 at 17:38:07 UTC from IEEE Xplore. Restrictions apply.

is also a research candidate in the cognitive robotics group of the Oxford Robotics Institute. He obtained a master's degree from Carnegie Mellon University and has worked for IBM Research as a research intern. Workshop co-organizing experiences include an explainability workshop at CHI, multiple workshops on AI for autonomous driving at NeurIPS and IJCAI, and volunteering for the Black in AI workshop at NeurIPS.

Seyun Kim is a Ph.D. student at Carnegie Mellon University in the Human-Computer Interaction Institute, Social AI Group. Her research goals are to explore fairness and transparency in algorithmic systems as well as mitigating biases in these systems. She obtained a master's degree from Cornell University at the Robots in Groups Lab focusing on group cohesion in human-robot teams.

Martim Brandao is a Post-Doctoral Research Associate at King's College London, whose research is related to explainability and fairness in planning and robotics methods. Martim has previously co-organized workshops on bias, fairness and ethics of robotics at ICRA and ARSO.

Min Kyung Lee is an Assistant Professor in the School of Information at the University of Texas at Austin. Dr. Lee's research examines the social implications of algorithms' emerging roles in management and governance in society, looking at the impacts of algorithmic management on workers as well as public perceptions of algorithmic fairness. She has also proposed participatory frameworks for designing algorithms with stakeholders, and conducted research on social robots and telepresence robots. Dr. Lee has served on the organization of FAccT, CHI, RSS and HRI and co-organized various workshops on topics of transparency, explainability, social justice, participatory approaches, responsible AI and others. She is a Co-PI on a new NSF NRT grant, Convergent, Responsible, and Ethical Artificial Intelligence Training Experience for Roboticists.

Malte F Jung is an Associate Professor of Information Science at Cornell University. His research focuses on understanding how we can design robots with interpersonal dynamics in mind.

REFERENCES

[1] S. Benthall and B. D. Haynes, "Racial categories in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 289–298.

[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[3] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 3315–3323, 2016.

[4] M. Veale, M. Van Kleek, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.

[5] H. Cramer, J. Garcia-Gathright, S. Reddy, A. Springer, and R. Takeo Bouyer, "Translation, tracks & data: an algorithmic bias effort in practice," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.

[6] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[7] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Reinforcement Learning with Fairness Constraints for Resource Distribution in Human-Robot Teams," Tech. Rep.

[8] M. L. Chang, G. Trafton, J. M. McCurry, and A. L. Thomaz, "Unfair! perceptions of fairness in human-robot teams," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 905–912.

[9] M. F. Jung, D. DiFranzo, B. Stoll, S. Shen, A. Lawrence, and H. Claure, "Robot assisted tower construction-a resource distribution task to study human-robot collaboration and interaction with groups of people," *arXiv preprint arXiv:1812.09548*, 2018.

[10] M. Brandao, M. Jirotka, H. Webb, and P. Luff, "Fair navigation planning: a resource for characterizing and designing fairness in mobile robots," *Artificial Intelligence*, vol. 282, p. 103259, 2020.

[11] S. K. Ötting, S. Gopinathan, G. W. Maier, and J. J. Steil, "Why criteria of decision fairness should be considered in robot design," 2017.

[12] J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. Porter, and K. Y. Ng, "Justice at the millennium: a meta-analytic review of 25 years of organizational justice research." *Journal of applied psychology*, vol. 86, no. 3, p. 425, 2001.

[13] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Fairness and the assumptions of economics," *Journal of business*, pp. S285–S300, 1986.

[14] W. M. Alves and P. H. Rossi, "Who should get what? fairness judgments of the distribution of earnings," *American journal of Sociology*, vol. 84, no. 3, pp. 541–564, 1978.

[15] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[16] M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish, "Working with machines: The impact of algorithmic and data-driven management on human workers," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 1603–1612.

[17] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, p. 2053951718756684, 2018.

[18] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A qualitative exploration of perceptions of algorithmic fairness," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.

[19] M. F. Jung, D. DiFranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–23, 2020.

[20] M. L. Chang, Z. Pope, E. S. Short, and A. L. Thomaz, "Defining fairness in human-robot teams," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1251–1258.

[21] M. Brandao, "Socially fair coverage: The fairness problem in coverage planning and a new anytime-fair method."

[22] M. Vázquez, A. Steinfeld, and S. E. Hudson, "Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 36–43.

[23] A. Addison, C. Bartneck, and K. Yogeeswaran, "Robots can be more than black and white: examining racial bias towards robots," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 493–498.

[24] M. Mansouri, B. Martim, and A. Saffiotti, "Bias-sensitizing robot behaviours: A quest for avoiding harmful bias and discrimination by robots (against-19)," 2019.

[25] M. Mansouri, B. Martim, and M. Magnusson, "Against robot dystopias: Thinking through the ethical, legal, and societal issues of robotics and automation (against-20)," 2019.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on June 30,2023 at 17:38:07 UTC from IEEE Xplore. Restrictions apply.