Human-Centered Approaches to Fair and Responsible AI

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

Nina Grgić-Hlača
Max Planck Institute for Software Systems
Saarbrücken, Germany
nghlaca@mpi-sws.org

Michael Carl Tschantz
International Computer Science Institute
Berkeley, CA USA
mct@icsi.berkeley.edu

Reuben Binns
University of Oxford
Oxford, UK
reuben.binns@cs.ox.ac.uk

Adrian Weller
University of Cambridge
Cambridge, UK
aw665@cam.ac.uk

Michelle Carney
Google AI
Mountain View, CA USA
michellecarney@google.com

Kori Inkpen
Microsoft Research
Redmond, WA USA
kori@microsoft.com

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3375158

Abstract

As AI changes the way decisions are made in organizations and governments, it is ever more important to ensure that these systems work according to values that diverse users and groups find important. Researchers have proposed numerous algorithmic techniques to formalize statistical fairness notions, but emerging work suggests that AI systems must account for the real-world contexts in which they will be embedded in order to actually work fairly. These findings call for an expanded research focus beyond statistical fairness to that which includes fundamental understandings of human uses and the social impacts of AI systems, a theme central to the HCI community. The HCI community can contribute novel understandings, methods, and techniques for incorporating human values and cultural norms into AI systems; address human biases in developing and using AI; and empower individual users and society to audit and control AI systems. Our goal is to bring together academic and industry researchers in the fields of HCI, ML and AI, and the social sciences to devise a cross-disciplinary research agenda for fair and responsible AI systems. This workshop will build on previous algorithmic fairness workshops at AI and ML conferences, map research and design opportunities for future innovations, and disseminate them in each community.

Author Keywords
Algorithmic fairness, accountability, ethics

CCS Concepts
•Human-centered computing → Human computer interaction (HCI);

Background
Advances in artificial intelligence (AI) have changed the way decisions are made in organizations, governments, and everyday life. Computational algorithms determine how information is distributed online [11], how organizations and work platforms assign and evaluate work and hire workers [22, 23], and how government programs allocate resources in cities and make judicial decisions [25]. It is crucial to ensure that these systems work according to values that diverse users and groups find important.

To address this issue, researchers have proposed numerous algorithmic techniques to formalize statistical fairness notions [8]. However, emerging work suggests that AI systems must account for the real-world contexts in which they will be embedded in order to actually work fairly. Fairness is often context-dependent, and many digital systems need fairness notions that go beyond nondiscrimination [15], which has been the main focus of statistical fairness research. Even when algorithms are fair in terms of the statistical outcomes, they can be perceived as unfair if the decision-making process is unfair or the AI does not respect community values and norms [18, 28]. Multiple stakeholders may have goals that conflict with each other and the AI system [20], so informed trade-off decisions must be made in an acceptable, fair way. Finally, human biases and organizational dynamics can lead to unfair decision outcomes [12, 13] or ultimately prevent the adoption of the AI system [26].

These findings call for an expanded research focus beyond statistical fairness that includes fundamental understandings of human uses and social impacts of AI systems, a theme central to the human–computer interaction (HCI) community. The HCI community can contribute novel understandings, methods, and techniques for incorporating human values and cultural norms into AI systems; address human biases in developing and using AI; and empower individual users and society to audit and control AI systems.

The goal of our workshop is to set up a human-centered research agenda for fair and responsible AI systems by bringing together academic and industry researchers and practitioners in the fields of HCI, machine learning (ML) and AI, and the social sciences. Several algorithmic fairness workshops have been organized at AI and ML conferences, but none have yet been held at HCI-oriented conferences. By building on previous workshops and organizing one at CHI 2020, we will leverage the expertise of the HCI community and set up a cross-disciplinary agenda for future research and design innovation, and disseminate it in each community. Particularly, we will seek to achieve three fundamental outcomes:

1) Synthesis of emerging research discoveries and methods. An emerging line of work seeks to systematically study human perceptions of algorithmic fairness [17, 24, 14], explain algorithmic decisions to promote trust and a sense of fairness [7, 10, 19], understand human use of algorithmic decisions [13, 12, 23], and develop methods to incorporate them into AI design [21, 29]. How can we map the current research landscape to identify gaps and opportunities for fruitful future research?

2) Tighter collaboration between HCI and AI/ML. How do we promote closer collaboration between HCI and AI/ML researchers to create novel approaches to fair AI in the real

world? For example, can algorithms predict when humans will be biased and intervene, and vice versa?

3) Design guidelines for fair and responsible AI. Existing fairness AI toolkits aim to support algorithm developers [1], and existing human-AI interaction guidelines mainly focus on usability and experience [5, 6]. Can we create design guidelines for HCI and user experience (UX) practitioners and educators to design fair and responsible AI?

Organizers

The organizing team consists of academic and industry researchers and practitioners in the field of HCI, UX, ML, privacy and security, and social sciences. Our organizers have been actively conducting research on the topic of fairness in AI systems and leading industry projects that incorporate AI into digital products. Many of our team members have organized workshops on human-centered AI and ML at CHI [16], CSCW [27], ICML [2], NeurIPS [4] and Interact [3]. The cross-disciplinary nature of our team will allow the organizers to attract workshop attendees from diverse backgrounds, stimulate interdisciplinary discussion during the workshop, and disseminate the results in each of their communities, spanning both academia and industry.

Min Kyung Lee will be the main contact person. She is an assistant professor in the School of Information at the University of Texas at Austin. She is an HCI researcher and draws from computer science, social science, and design theory and methods. She has proposed a participatory framework that empowers community members to design matching algorithms for their own communities [21]. She has also conducted one of the first studies to investigate the impacts of algorithmic management on workers [17, 22] as well as public perceptions of algorithmic fairness [18, 19].

Nina Grgić-Hlača is a PhD student at the Max Planck Institute for Software Systems. Prior to joining the Max Planck Institute, she received an MA in Informatics and Philosophy from the University of Zagreb, Croatia. Her research interests include algorithmic fairness, human-centered ML and social computing. She is particularly interested in studying the interaction between humans and ML-based systems in the context of fair decision-making.

Michael Carl Tschantz received a Ph.D. from Carnegie Mellon University in 2012 and a Sc.B. from Brown University in 2005, both in Computer Science. Before becoming a researcher at the International Computer Science Institute in 2014, he did two years of postdoctoral research at the University of California, Berkeley. He uses the models of AI and statistics to solve the problems of privacy and security. His interests also include experimental design, formal methods, and logics. His current research includes automating information flow experiments, circumventing censorship, and securing ML. His dissertation formalized and operationalized what it means to use information for a purpose.

Reuben Binns is a Research Fellow at the Information Commissioner's Office, addressing AI/ML and data protection. He conducts research and teaches part time at the Department of Computer Science at the University of Oxford, where he is currently funded by the EPSRC's PETRAS Internet of Things Research Hub project 'Respectful Things in Private Spaces'. His recent work has focused on two strands: human factors of third-party tracking on the web, mobile and Internet-of-Things devices; and transparency, fairness and accountability in profiling and ML.

Adrian Weller is a principal research fellow in ML at the University of Cambridge and Programme Director for AI at The Alan Turing Institute (UK national institute for data science and AI), where he is also a Turing Fellow leading work

on safe and ethical AI. He is a principal research fellow at the Leverhulme Centre for the Future of Intelligence leading their Trust and Transparency project. His interests span AI, its commercial applications and helping to ensure beneficial outcomes for society. He serves on several boards including the Centre for Data Ethics and Innovation. Previously, Adrian held senior roles in finance.

Michelle Carney is a Computational Neuroscientist turned UX Practitioner. Her practice focuses on the intersection of Data Science/ML/AI and UX. Currently a Senior UX Researcher on Google's AIUX Team, Michelle's projects focus on combining ML and UX for projects at Google Research, from using data science techniques to inform UX as well as leading UXR to inform the direction of ML. Outside of work, Michelle organizes the ML and UX Meetup (https://www.meetup.com/MLUXSF/) and teaches at the Stanford d.school as a Lecturer for the Designing ML course and various workshops.

Kori Inkpen is a Senior Principal Researcher at Microsoft Research. Her current research is focused on Human+AI Collaboration to improve human decision making, particularly in high-stakes contexts. A key component of this research is understanding Bias and Fairness in Human+AI systems, and exploring ways to mitigate these biases. Previously, her work explored the use of video to transform the way we interact with friends, families, and colleagues.

Website

Our workshop website (http://fair-ai.owlstown.com) will explain the goal and logistics of the workshop, including the call for participation (CFP), instructions for the position papers, workshop schedule, and organizers' backgrounds. We will post accepted position papers, other readings related to the workshop theme and updates and news after the workshop.

Pre-Workshop Plans

We will recruit participants by sending out a call for participation via HCI and algorithmic fairness-related mailing lists such as CHI-announcements and FATML as well as participant lists from the workshops that we have previously organized. We will also post the CFP in UX designers' working groups and meet-ups in order to reach industry practitioners. The accepted position papers will be shared on the workshop website for all participants. The organizers will also derive high-level categories from the work presented in the position papers and a set of application areas for the workshop activities, which are described in the next section.

Workshop Structure

The workshop will last one day and include presentations, group work sessions, and discussions (Table 1). We will limit the participants to 20 excluding the organizers.

Introduction

The organizers will welcome the participants and set the stage for the workshop by explaining the workshop goals.

Mapping the Landscape of Research and Practice

The first half of the workshop will focus on having the participants collectively map the current landscape of research and practice for fair and responsible AI systems. First, each participant will give a "madness" style, four-minute presentation on their position papers. The goal is to familiarize all participants with cutting-edge discoveries and techniques related to the workshop topic both in academia and industry, sharing opportunities and challenges with each other. The presentations will occur over two one-hour sessions with a coffee break in between. The participants will be provided with pens and post-its notes and encouraged to write down themes that they observe during the sessions. The prompts for note-taking will ask about the presentation's

design implications, whether its findings are consistent or inconsistent with other presentations, what assumptions underline the work, etc. We will provide high-level categories on multiple poster-size sheets of paper and put them on a wall, so that participants can post their notes during the break. After the presentation sessions, we will collectively categorize the post-it notes and synthesize emerging themes. Each organizer will lead the process for each category, and participants will be asked to work on the category that interests them most. The synthesis session will conclude with a brief presentation from each group.

Keynote Presentation
The lunch break will be followed by a 30-minute keynote presentation and a 10-minute Q&A. We will invite an expert at the forefront of human-centered research on transparent and fair ML. Tentative speakers include Ece Kamar, Krishna Gummadi, Rich Caruana, and Alison Gopnik.

Identifying Opportunities and Guidelines
The second half of the workshop will be in breakout group sessions. The goal is to build on the themes learned in the morning, and work on identifying future research opportunities and devising design guidelines for practice. In order to ground these activities on concrete real world contexts, we will offer a set of applications for participants to work on. The example applications will include online content moderation algorithms, voice agents that learn and assist users' daily activities, algorithms that assess credit scores and health insurance claims, judicial algorithms that make recidivism and criminal sentencing decisions, and algorithmic systems that allocate resources in cities, such as school assignment and public transportation scheduling. We will provide personas of stakeholders, types of data and algorithms, and cultural and geographical contexts.

We will conduct these afternoon activities in the convention center rooftop patio, with different work stations in the patio for each of the applications. Each station will be run by one of the organizers, and participants will pick one application to work on. In the first group session, participants will be provided worksheets that contain prompts that guide the group activities. Example prompts include possible utopian and dystopian scenarios, their implications on the stakeholders, and the applicability of the themes drawn from the morning sessions, as well as remaining open questions and concerns not covered by those themes. In the second group session, participants will work on organizing future research questions and potential studies and structure design guidelines for practice. In the last session, each group will give a 7 minute presentation on their work.

Wrap-Up and Discussion of Next Steps
We will wrap up with a discussion of next steps. We will collect the topics emerged from the workshop, and the names of participants who are interested in continuing to work on them. We will also survey potential outlets for future publications, guidebooks, workshops, and special issues of journals on the topic.

Post Workshop Plans

The workshop website will be kept up to date after the workshop in order to serve as a repository of resources for research and practice of fair AI systems. We will post the results of the workshop and subsequent publications and handbooks. We will also create working groups based on the information collected at the end of the workshop. One tangible goal of the working groups will be to write a review and vision-oriented article (like [9]) for academic venues such as ACM FAT*. Another tangible goal of the working group will be to write a guideline for industry practitioners and post it on the workshop website to disseminate through UX practitioners' working groups.

Time	Activity
9:00-9:10	Introduction
9:10-10:10	Position paper presentations I
10:10-10:30	Coffee break
10:30-11:30	Position paper presentations II
11:30-12:30	Emerging theme synthesis
12:30-13:40	Lunch
13:40-14:20	Keynote presentation
14:20-15:20	Group work session I
15:20-15:40	Coffee break
15:40-16:40	Group work session II
16:40-17:15	Group presentations
17:15-17:30	Wrap-up and next steps
After 17:30	(Optional) Workshop dinner

Table 1: Workshop Schedule

Call for Participation

As AI changes the way decisions are made in organizations and governments, it is ever more important to ensure that these systems work according to values that diverse users and groups find important. Researchers have proposed numerous algorithmic techniques to formalize statistical fairness notions, but emerging work suggests that AI systems must account for the real-world contexts in which they will be embedded in order to actually work fairly. These findings call for an expanded research focus beyond statistical fairness that includes fundamental understandings of human use and the social impact of AI systems, a theme central to the HCI community. This workshop aims to bring together a diverse group of researchers and practitioners to develop a cross-disciplinary agenda on creating fair and responsible AI systems. To participate, submit a 2-4 page paper in CHI extended abstract format to minkyung.lee@austin.utexas.edu. Potential topics include:

- Human biases in human-in-the-loop decisions
- Human perceptions of algorithmic fairness
- Human-centered evaluation of fair ML models
- Explanations & transparency of algorithmic decisions
- Methods for stakeholder participation in AI design
- Decision-support system design
- Algorithm auditing techniques
- Ethics of AI
- Sociocultural studies of AI in practice

Position papers will be reviewed by the organizers and evaluated based on their quality, novelty, and fit with the workshop theme. At least one author of an accepted paper must attend the workshop and all participants must register for both the workshop and for at least one day of the conference. Important dates:

- Position paper deadline: February 11, 2020
- Notification: February 28, 2020
- Workshop at CHI2020: April 25/26, 2020

More Information:http://fair-ai.owlstown.com

REFERENCES

[1] 2019. IBM AI Fairness 360 Open Source Toolkit. (2019). https://aif360.mybluemix.net.

[2] 2019. ICML 2019 Workshop on Human In the Loop Learning. (2019). https://sites.google.com/view/hill2019/home.

[3] 2019. INTERACT 2019 Workshop on Humans in the Loop – Bringing AI & HCI Together. (2019). https://koriinkpen.github.io/Interact2019_AI-HCI/.

[4] 2019. NeurIPS 2019 Workshop on Human-centric Machine Learning. (2019). https://sites.google.com/view/hcml-2019.

[5] 2019. People+AI Guidebook. (2019). https://pair.withgoogle.com.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, and others. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[7] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[8] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[9] John Danaher, Michael J Hogan, Chris Noone, Rónán Kennedy, Anthony Behan, Aisling De Paor, Heike Felzmann, Muki Haklay, Su-Ming Khoo, John Morison, and others. 2017. Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society* 4, 2 (2017).

[10] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 275–285.

[11] Tarleton Gillespie. 2010. The politics of "platforms". *New Media & Society* 12, 3 (2010), 347–364.

[12] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.

[13] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 178.

[14] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, 903–912.

[15] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 600.

[16] Kori Inkpen, Stevie Chancellor, Munmun De Choudhury, Michael Veale, and Eric PS Baumer. 2019. Where is the Human?: Bridging the Gap Between AI and HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, W09.

[17] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 1–16.

[18] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1035–1048.

[19] Min Kyung Lee, Anuraag Jain, Hae Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM : Human-Computer Interaction* 3, CSCW (2019), Article 182, 26 pages.

[20] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3365–3376.

[21] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel Procaccia. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM : Human-Computer Interaction* 3, CSCW (November 2019), Article 181, 34 pages.

[22] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. ACM, 1603–1612.

[23] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. (2019).

[24] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction-CSCW* 3 (2019), 28.

[25] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. 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*. ACM, 440.

[26] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, and Jason Schultz. 2018. AI Now Report 2018. (2018).

[27] Christine T Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R Brubaker. 2018. The Changing Contours of Participation in Data-driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 377–384.

[28] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Paper 656.

[29] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), article 194.