Human-Centered Responsible Artificial Intelligence: Current & Future Trends

Mohammad Tahaei
Nokia Bell Labs
Cambridge, UK
mohammad.tahaei@nokia-bell-labs.com

Marios Constantinides
Nokia Bell Labs
Cambridge, UK
marios.constantinides@nokia-bell-labs.com

Daniele Quercia
Nokia Bell Labs
Cambridge, UK
daniele.quercia@nokia-bell-labs.com

Sean Kennedy
Nokia Bell Labs
Ottawa, Canada
sean.kennedy@nokia-bell-labs.com

Michael Muller
IBM Research AI
Cambridge, MA, USA
michael_muller@us.ibm.com

Simone Stumpf
University of Glasgow
Glasgow, UK
simone.stumpf@glasgow.ac.uk

Q. Vera Liao
Microsoft Research
Montreal, Canada
veraliao@microsoft.com

Ricardo Baeza-Yates
EAI, Northeastern University
CA, USA
rbaeza@acm.org

Lora Aroyo
Google
NY, USA
l.m.aroyo@gmail.com

Jess Holbrook
Meta
USA
jess.holbrook@gmail.com

Ewa Luger
University of Edinburgh
Edinburgh, UK
ewa.luger@ed.ac.uk

Michael Madaio
Google
USA
madaiom@google.com

Ilana Golbin Blumenfeld
PwC
Los Angeles, CA, USA
ilana.a.golbin@pwc.com

Maria De-Arteaga
University of Texas at Austin
TX, USA
dearteaga@mccombs.utexas.edu

Jessica Vitak
University of Maryland, College Park
MD, USA
jvitak@umd.edu

Alexandra Olteanu
Microsoft Research
Montreal, Canada
alexandra.olteanu@microsoft.com

ABSTRACT

In recent years, the CHI community has seen significant growth in research on *Human-Centered Responsible Artificial Intelligence*. While different research communities may use different terminology to discuss similar topics, all of this work is ultimately aimed at developing AI that benefits humanity while being grounded in human rights and ethics, and reducing the potential harms of AI. In this special interest group, we aim to bring together researchers from academia and industry interested in these topics to map current and future research trends to advance this important area of research by fostering collaboration and sharing ideas.

CCS CONCEPTS

• **Social and professional topics**; • **Human-centered computing**; • **Theory of computation**; • **Information systems**; • **Software and its engineering**; • **Security and privacy**;

KEYWORDS

human-centered AI, responsible AI, AI ethics

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3583178

ACM Reference Format:
Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q. Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, Ewa Luger, Michael Madaio, Ilana Golbin Blumenfeld, Maria De-Arteaga, Jessica Vitak, and Alexandra Olteanu. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544549.3583178

1 MOTIVATION & BACKGROUND

Human-Centered Responsible Artificial Intelligence (HCR-AI)[1] aims to bring people and their values into the design and development of AI systems, which can contribute to building systems that benefit people and societies, as well as preventing and mitigating potential harms. Despite a long history of the importance of the human factor in AI systems [12, 31], there has been a growing awareness of its importance within the CHI community in the past few years [32]. Searching the ACM Digital Library within CHI proceedings shows the following results (Figure 1):[2] "human-centered AI" results in 41 records since 2019 and "responsible AI" results in 32 records since 2020. Below, we highlight a few examples of these studies, which are relevant to the topic of the Special Interest Group (SIG), noting that this is not an exhaustive list and is only to show the breadth and depth of the existing work:

Ethics in AI involve socio-cultural and technical factors, spanning a range of responsible AI values (including but not limited to transparency, fairness, explainability, accountability, autonomy, sustainability, and trust) [20]. However, different stakeholders, including the general population and AI practitioners, may perceive and prioritize these values differently. For example, a representative sample of the U.S. population was more likely to value safety, privacy, and performance. In contrast, practitioners were more likely to prioritize fairness, dignity, and inclusiveness [19]. Or, certain historically exploited groups may weigh privacy or non-participation more highly than groups with lower risk [13, 26].

Aligned with responsible AI are calls to make AI more human-centric. In particular, there is an emphasis on the challenges of AI integration into socio-technical processes to preserve human autonomy and control, as well as the impacts of AI systems deployment and applications on society, organizations, and individuals [4]. On this strand of research, understanding *socio-technical* and *environmental* factors can help surface why and how an AI system may become human-centered [8, 24, 30]. For example, even for an AI for which there might be broader consensus on its utility, such as the detection of diabetes using retina scans, there may well be barriers to becoming useful for its intended users, including due to not fitting well with the users' workflows (e.g., nurses) or the system requiring high-quality images that are not easy to produce, especially in locations with low resources where such technology can provide significant support to patients if done right [2].

Similarly, researchers have looked at individuals' expectations and understandings of AI. For example, when making an *ethical decision* (e.g., a hypothetical scenario for bringing down a terrorist drone to save lives), people may put more *capability trust* in an AI decision maker (i.e., capacity trustworthiness, being more capable), whereas they may put more *moral trust* in a human expert (i.e., being able to be morally trustworthy and make decisions that are aligned with moral values); in either case, decision made by a human or an AI, prior work has found that people often see the human as partly responsible, be it the decision maker or the AI developer [33]—even though the outcomes of the developer may intentionally or unintentionally limit the span of action of the decision-maker [27]. Regarding moral dilemmas between AI and human decisions, people may not equally judge humans and machines [17]. These variations in perceptions may be rooted in (a) people judging humans by their intentions and machines by their outcomes, and (b) people assigning extreme intentions to humans and narrow intentions to machines, while they may excuse human actions more than machine actions in accidental scenarios [17]. Furthermore, people's perceived fairness and trust in an AI may change with the terminology used to describe it (e.g., an algorithm, computer program, or artificial intelligence), which could eventually impact the system's success and outcomes, especially when comparative research is done [21].

Another human aspect of AI systems is the people who work on these systems, such as annotators, engineers, and researchers. Data annotators are part of the workforce that produces the datasets used to train AI models. However, the workforce (sometimes referred to as *AI labor* [5] or *ghostworkers* [15]) behind the annotation task may have career aspirations that the current annotation companies do not support, or they may be poorly paid because of the push that comes from the recent development in AI that requires massive annotated datasets at low costs [14, 34]. Other researchers echo similar observations about AI labor by saying that "without the work and labor that were poured into the data annotation process, ML [Machine Learning] efforts are no more than sandcastles," [34] or "everyone wants to do the model work, not the data work," [29] a behavior that contributes to the creation of *data cascades*—which refer to compounding events causing adverse, downstream effects from data issues, resulting in technical debt.[3]

New tools and frameworks are now being proposed to help developers build more responsible AI systems (e.g., IBM's 360 suites on fairness and explainability [18, 25] and Fairlearn [3]), in addition to user-led approaches to algorithmic auditing to uncover potential harms of algorithmic systems [7]. Despite the growing interest in HCI research and user experience design for AI, developing responsible AI remains challenging; a mission involving cognitive, socio-technical, cultural, and design perspectives [16, 23, 24].

These are just a few examples from many studies that cover topics that have emerged within the past few years and are relevant to the SIG's scope. Besides CHI, the ACM Conference on Fairness, Accountability, and Transparency (*ACM FAccT*), established in 2018 [1], aims to bring "together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems" highlighting the importance of the research in HCR-AI. We aim to bring this community together in a 75-minute discussion and brainstorming session at CHI 2023.

[1] Different communities have adopted different terminologies to address related topics. We intentionally left the proposal and terminology open without emphasizing specific topics to attract participants from various backgrounds and interests. One reason to propose this SIG is to discuss various aspects of HCR-AI with researchers who can provide diverse perspectives.

[2] Using search within anywhere on the ACM Digital Library. Results are not mutually exclusive and include all types of materials (e.g., research papers, extended abstracts, panels, and invited talks). Filtering for only research papers results in 32 unique papers since 2020. We acknowledge this is not an exhaustive search and is only to show the growing body of research in CHI.

[3] In 1992, Ward Cunningham put forward the metaphor of technical debt to describe the build-up of cruft (deficiencies in internal quality) in software systems as debt accrual, similar to financial [6] or ethical debt [11] (i.e., "AI ethical debt is incurred when an agency opts to design, develop, deploy and use an AI solution without proactively identifying potential ethical concerns" [28]).

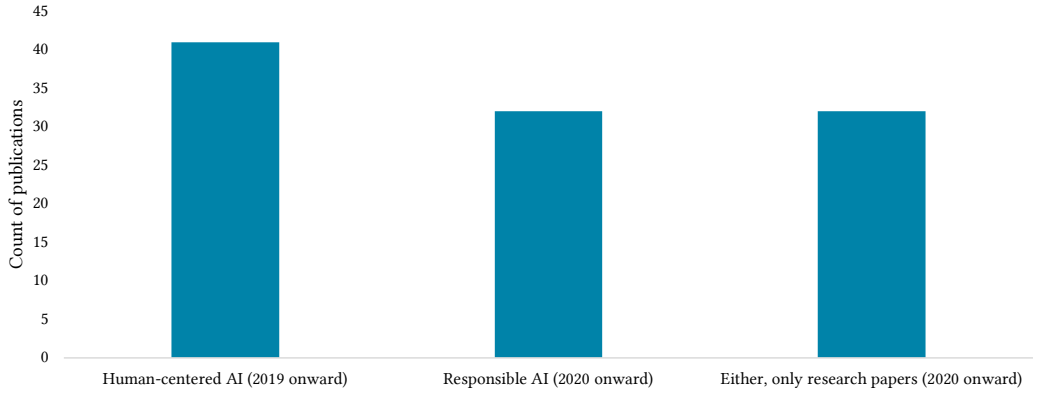

Figure 1: Counts of publications containing "human-centered AI" and "responsible AI" at CHI. The first three bars for human-centered AI and responsible AI are not mutually exclusive. They include all types of materials (e.g., research papers, extended abstracts, and invited talks). Filtering for only research papers results in 32 unique papers since 2020 (the last bar).

2 PROPOSAL & SIG'S GOAL

The SIG follows similar strands from past workshops at CHI 2020, 2021, and 2022 [9, 10, 22]. The topics discussed are evolving and growing (Figure 1); hence, a SIG at CHI 2023 would be timely. We believe a SIG dedicated to the HCR-AI at CHI 2023 will benefit the CHI community and help build and establish a broader network of researchers and provide a mapping and understanding of current and future trends in this area. Researchers in this area come from industry and academia from diverse disciplinary backgrounds (e.g., theoretical computer science, social computing, machine learning, human-computer interaction, and social science). Therefore, having them all in one hybrid physical-virtual room for 75 minutes would benefit the community and the attendees to brainstorm and generate a map of current and future trends in this area (activity diagramming). We propose to use online tools such as Miro and Slack to (a) create a record of the group's co-constructed knowledge; (b) serve as a persistent communication to others in the CHI community; and (c) enfranchise remote participants.

3 EXPECTED OUTCOMES & NEXT STEPS

We will share the Miro board with attendees and make it public to support future research in HCR-AI. We will also create a Slack channel for future communications. The SIG's primary goal is to create a sense of community among researchers in this area, from academia and industry, to establish collaborations. The SIG is an excellent opportunity to bring people with a shared interest in HCR-AI who also attend CHI to build this community.

After the SIG, we will organize virtual biannual meetings with the attendees to share their latest ideas and recent work, build a website to share outcomes created during the SIG, encourage attendees to apply for joint grants, and explore the possibility of creating a symposium similar to CHIWORK.

REFERENCES

[1] ACM. 2022. *ACM FAccT*. ACM. Retrieved November 2022 from https://facctconference.org

[2] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–12. https://doi.org/10.1145/3313831.3376718

[3] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[4] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. In *In the Navigating the Broader Impacts of AI Research Workshop at NeurIPS 2020.* https://www.microsoft.com/en-us/research/publication/overcoming-failures-of-imagination-in-ai-infused-system-development-and-deployment/

[5] Kate Crawford. 2021. *Atlas of AI.* Yale University Press.

[6] Ward Cunningham. 1992. The WyCash Portfolio Management System. In *Addendum to the Proceedings on Object-Oriented Programming Systems, Languages, and Applications (Addendum) (OOPSLA '92)*. ACM, 29–30. https://doi.org/10.1145/157709.157715

[7] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating Users' Strategies for Uncovering Harmful Algorithmic Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, Article 626, 19 pages. https://doi.org/10.1145/3491102.3517441

[8] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, 19 pages. https://doi.org/10.1145/3411764.3445188

[9] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. ACM, Article 94, 6 pages. https://doi.org/10.1145/3411763.3441342

[10] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, Article 109, 7 pages. https://doi.org/10.1145/3491101.3503727

[11] Casey Fiesler and Natalie Garrett. 2020. *Ethical Tech Starts with Addressing Ethical Debt.* Wired. Retrieved December 2022 from https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/

[12] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[13] Patricia Garcia, Tonia Sutherland, Marika Cifor, Anita Say Chan, Lauren Klein, Catherine D'Ignazio, and Niloufar Salehi. 2020. No: Critical Refusal as Feminist

Data Practice. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion)*. ACM, 199–202. https://doi.org/10.1145/3406865.3419014

[14] Sandy J. J. Gould. 2022. Consumption Experiences in the Research Process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, Article 326, 17 pages. https://doi.org/10.1145/3491102.3502001

[15] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books. https://ghostwork.info

[16] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019). https://doi.org/10.1126/scirobotics.aay7120

[17] César A Hidalgo, Diana Orghian, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines*. MIT Press. https://doi.org/10.7551/mitpress/13373.001.0001

[18] IBM. 2022. *AI Fairness 360*. IBM. Retrieved December 2022 from https://aif360.mybluemix.net

[19] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 310–323. https://doi.org/10.1145/3531146.3533097

[20] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[21] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 28 pages. https://doi.org/10.1145/3491102.3517527

[22] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered Approaches to Fair and Responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. ACM, 1–8. https://doi.org/10.1145/3334480.3375158

[23] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, 14 pages. https://doi.org/10.1145/3411764.3445570

[24] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. (2021). https://doi.org/10.48550/ARXIV.2110.10790

[25] Aleksandra Mojsilovic. 2019. *Introducing AI Explainability 360*. IBM. Retrieved December 2022 from https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

[26] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 19 pages. https://doi.org/10.1145/3491102.3517644

[27] Michael Muller and Justin Weisz. 2022. Extending a Human-AI Collaboration Framework with Dynamism and Sociality. In *2022 Symposium on Human-Computer Interaction for Work (CHIWORK 2022)*. ACM, Article 10, 12 pages. https://doi.org/10.1145/3533406.3533407

[28] Catherine Petrozzino. 2021. Who pays for ethical debt in AI? *AI and Ethics* 1, 3 (2021), 205–208. https://doi.org/10.1007/s43681-020-00030-3

[29] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, 15 pages. https://doi.org/10.1145/3411764.3445518

[30] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 26 (Oct 2020), 31 pages. https://doi.org/10.1145/3419764

[31] Lucy A. Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

[32] Mohammad Tahaei, Marios Constantinides, and Daniele Quercia. 2023. Toward Human-Centered Responsible Artificial Intelligence: A Review of CHI Research and Industry Toolkits. 9 pages. https://doi.org/10.48550/ARXIV.2302.05284

[33] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 17 pages. https://doi.org/10.1145/3491102.3517732

[34] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 16 pages. https://doi.org/10.1145/3491102.3502121