Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI

Upol Ehsan
Georgia Institute of Technology
Atlanta, GA, USA
ehsanu@gatech.edu

Philipp Wintersberger
TU Wien
Vienna, Austria
philipp.wintersberger@carissma.eu

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

Elizabeth Anne Watkins
Princeton University
Princeton, New Jersey, USA
ew4582@princeton.edu

Carina Manger
Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, Germany
Carina.Manger@carissma.eue

Hal Daumé III
University of Maryland, College Park, MD, USA
Microsoft Research, NY, USA
hal3@umd.edu

Andreas Riener
Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, Germany
andreas.riener@thi.de

Mark O. Riedl
Georgia Institute of Technology
Atlanta, GA, USA
riedl@cc.gatech.edu

ABSTRACT

Explainability of AI systems is crucial to hold them accountable because they are increasingly becoming consequential in our lives by powering high-stakes decisions in domains like healthcare and law. When it comes to Explainable AI (XAI), understanding *who* interacts with the black-box of AI is just as important as "opening" it, if not more. Yet the discourse of XAI has been predominantly centered around the black-box, suffering from deficiencies in meeting user needs and exacerbating issues of algorithmic opacity. To address these issues, researchers have called for human-centered approaches to XAI. *In this second CHI workshop on Human-centered XAI (HCXAI)*, we build on the success of the first installment from CHI 2021 to expand the conversation around XAI. We chart the domain and shape the HCXAI discourse with reflective discussions from diverse stakeholders. The goal of the second installment is to go beyond the black box and examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological, and technical levels. Encouraging holistic (historical, sociological, and technical) approaches, we put an emphasis on "operationalizing", aiming to produce actionable frameworks, transferable evaluation methods, concrete design guidelines, and articulate a coordinated research agenda for XAI.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence.

CHI '22, April 30–May 06, 2022, New Orleans, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3503727

KEYWORDS

Explainable Artificial Intelligence, Interpretable Machine Learning, Interpretability, Artificial Intelligence, Responsible AI, Trust in Automation, Algorithmic Fairness

ACM Reference Format:
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 *CHI '22: ACM Conference on Human Factors in Computing Systems, April 30–May 06, 2022, New Orleans, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3503727

1 INTRODUCTION

As AI-driven systems increasingly power high-stakes decision-making in public domains such as healthcare [16], finance [22], and criminal justice [26], their explainability is critical for end-users to take informed and accountable actions [28]. Issues concerning explainability lie at the heart of Explainable AI (XAI), a research area that aims to provide human-understandable information for the system's behavior and processes [3, 9]. While explainability in AI is not new [25, 27], recent proliferation of Deep learning based approaches—models of which are hard to understand even by experts—has ushered a a remarkable growth in techniques to "open" the black-box of AI [13].

When it comes to making AI systems explainable, understanding *who* opens and interacts with the black-box of AI matters just as much, if not more, as opening it. Explainability is as much a human factors problem as it is a technical one. Implicit in Explainable AI is the question: "explainable to whom?" [8]. The *who* governs the most effective way of describing the *why* behind the decisions. In the algorithmic-centered discourse, this human side of making AI systems explainable remains largely unexplored [8]. Moreover, algorithmic transparency alone does not encapsulate the sociotechnical notion of AI explainability [6]. It is is a multi-stakeholder problem that spans the entire AI lifecycle. Explainability is as needed

just as much by creators (e.g., engineers, managers, administrators) of AI systems but also by end-users, if not more - and these stakeholders all bring disparate and distinct goals to the task of explanation [19]. According to the European Parliament, people have a right to explainability whenever an AI system significantly impacts their lives [18]. To foster fair, ethical, and accountable AI, we need to go beyond opening the black-box and towards the human factors of XAI. Given XAI is as much HCI's problem as it is AI's, CHI is an appropriate venue to ensure the human side of XAI is appropriately addressed.

In this ***second*** **installment** of the Human-centered Explainable AI (HCXAI) workshop at CHI, we build on the groundwork of the first workshop [10] that was attended by over 100 participants from more than 14 countries. There is a critical need to continue and expand the conversation around HCXAI. Our first workshop attracted stakeholders from diverse domains such as public policy, machine learning, data visualization, psychology, law, and design. Receiving more than 50 submissions (24 accepted), we had a critically constructive dialogue around different human-centered perspectives in XAI. Notably, participants discussed (1) tensions around evaluating XAI systems, (2) unpacking what "human-centered" means in different contexts, and (3) actionable design guidelines for accountability in AI systems. Beyond a constructive dialogue, using our online space in Discord, we were also able to build a supportive community. Not only has that community conversation led to research collaboration amongst participants, a special ACM TiiS journal issue on HCXAI has just been accepted.

In this second workshop, we aim to offer a junction point for cross-disciplinary stakeholders to extend the conversation on the human-side of Explainable AI at the conceptual, methodological, and technical levels. The goals are to (1) extend the critically constructive dialogue around *operationalizing* human-centered perspectives in XAI and (2) further expand and foster a supportive the HCXAI community of diverse stakeholders.

2 FROM XAI TO HUMAN-CENTERED XAI: BACKGROUND AND OPPORTUNITIES

Systematically approaching XAI from a human-centered perspective requires to establish a common ground for discourse. From the perspective of "Social Construction of Technology" (SCOT) [4], relevant social groups (i.e., different stakeholders such as researchers, policy makers, etc.) draw on explainability's *interpretive flexibility* (i.e., the flexibility to support multiple concurrent diverging interpretations), which results in fluidity regarding the relevant constructs. Consequently, terms such as explainability, interpretability, or transparency, have been used interchangeably within different communities [1, 3, 24]. Some have defined explainability as an AI systems' decisions being *easy to understand* by people [3, 13], and the term is often viewed more broadly than transparency or interpretable models [21]. This is illustrated by a growing area within the XAI community, which addresses *post-hoc explanations* [9] that communicate an opaque model's decisions in a way that's accessible for end users [21], rather than exactly describing how the model works. Thus, a suitable "operationalization" requires contextually situating ambiguities among the involved research communities regarding definitions, concepts, or evaluation methods.

There is a tendency to design AI explanations *as if* people like the XAI developers are going to use their systems, earning an infamous reputation of "inmates running the asylum" [23]. In contrast, our workshop aims to bring user-centered perspectives to the forefront to support the diverse needs of involved stakeholders like including data scientists, decision makers, regulatory agencies, end-users. Each of these stakeholder groups differ when it comes to the interpretation of and reaction to explanations, which depends on multiple social and individual characteristics like professional and educational backgrounds [7, 12].

In HCXAI, we not only ask about for *whom* an explanation is created, but also *why* [8], since explanations are requested for a wide range of purposes such as *trustworthiness, causality, fairness, accessibility, interactivity, or privacy-awareness* [3]. Understanding *who* and *why* influences how and which data is collected. Providing the example of an automated vehicle, it is clear that engineers, sales agencies, policy makers, drivers, etc. require different forms of explanations. Liao et al. [20] have summarized such user goals and argue that XAI must aim at getting better insights into decision-making to appropriately evaluate AI's capabilities, adapt its usage, and improve its performance.

Beside the *why* and the *who*, the *"where"* and the *"when"* (i.e., the application domain or context) also play important roles. For example, recent work has introduced XAI features into model development tools [12], AI-assisted decision-support tools [29], model fairness evaluation [5], and more. The "when" has been paid little attention in XAI research so far. Typically, explanations are provided together (i.e., simultaneously or shortly after) with a corresponding AI decision. General principles such as whether to provide explanations *in-situ* versus for retrospective analysis of AI decisions do not exist – in some time-critical domains interacting with immediately presented explanations may even be impossible. What if, for instance, a future physician cooperating with an AI diagnosis system would inspect explanations both during the stressful times of the busy workday (i.e., when AI decisions are taken), and during other times that allow more deeply reflecting about the system's behavior? In a study of an ML model assessing risk of disease, physicians' receptivity to ML outputs changed depending on their stress levels and fatigue [11]. Such questions of timing will need to move into focus, since the effectiveness of explanations may depend on factors of timing as well.

Additionally, we argue it is important to include other communities so as not to reinvent the wheel. The human factors community is already engaged in questions around automation transparency and trustworthiness, and provided both potential design solutions and evaluation methods [15, 17]. Additionally, methodological contributions are necessary to evaluate XAI systems and stakeholder perceptions in all phases of the (human-centered) design process. Arriata et al. [3] emphasize that defining metrics should *"be approached by the community as a whole"* to allow evaluating and comparing XAI approaches under different application contexts, models and purpose. An in-depth review by Mohseni et al. [24] has identified different types of evaluation methods used in the literature, including both subjective (interviews, surveys, self-reports, etc.) and objective (task performance, user prediction of model

output, compliance, etc.) assessment techniques. A so-called "explanation satisfaction scale" has been proposed by Hoffman et al. [14]. The scale includes 8 items for relevant factors like understanding, satisfaction, accuracy, trust, etc. However, considering all the issues describe above, we believe sophisticated evaluation methods of the future will demand to better represent the *who, why, where, and when* than the tools already available.

All these developments are just the beginning of systematically understanding how real-world AI systems can be developed that are socially situated and culturally embedded. A critical perspective is essential to address intellectual blind spots and develop HCXAI in a systematic manner. The goal is *not* to impose a normativity but systematically articulate the different *interpretive flexibilities* of each *relevant social groups* in XAI. This allows us to actionably advance the XAI discourse from conceptual, methodological, and technical angles.

3 GOALS OF THE WORKSHOP

By facilitating a junction of diverse perspectives from relevant stakeholders in XAI, we want to achieve clarity in charting the future of XAI from historical, sociological, and technological perspectives. Building on the first workshop's results, the **goals** are to (1) extend the critically constructive dialogue around *operationalizing* human-centered perspectives in XAI and (2) further expand and foster a supportive the HCXAI community of diverse stakeholders. Bridging works from researchers, designers, and practitioners from the fields of XAI, HCI, psychology, machine learning, and social sciences, we want to *re-center our focus on the human*. Going beyond *just* opening the black-box, we aim to discuss and operationalize Human-centered XAI at the conceptual, methodological, and technical levels. Operationalization can include aspects such as frameworks, transferable evaluation methods, and actionable design guidelines.

Thus, we are interested in a wide range of topics, from sociotechnical aspects of XAI to human-centered evaluation techniques to responsible use of XAI. We are especially interested in the discourse around one or more of the questions discussed below: *who* (e.g., clarifying *who* the human is in XAI), *why* (e.g., how individual and social factors influence explainability goals), and *where* (e.g., explainability differences across application domains).

Beyond these, we also want to continue the conversation around *issues identified by participants in the first HCXAI workshop* [10] around topics like weaponizing AI explanations (e.g., inducing over-trust in AI) harmful effects, designing for accountability, and avoiding "ethics washing" in XAI. With an effort towards equitable conversations, we welcome participation from the Global South and from stakeholders whose voices are under-represented in the dominant XAI discourse. The following list of guiding questions is *not* an exhaustive one; rather, it is provided as sources of inspiration:

- How might we chart the landscape of different *'whos'* (relevant stakeholders) in XAI and their respective explainability needs?
- What user goals should XAI aim to support, for whom, and why?
- How can we address value tensions amongst stakeholders in XAI?
- How do user characteristics (e.g., educational background, profession) impact needs around explainability?
- Where, or in what categories of AI applications, should we prioritize our XAI efforts on?
- How might we develop transferable evaluation methods for XAI? What key constructs need to be considered?
- Given the contextual nature of explanations, what are the potential pitfalls of evaluation metrics standardization? How might we take into account the who, why, and where in the evaluation methods?
- How might we stop AI explanations from being weaponized (e.g., inducing dark patterns)?
- Not all harms are intentional. How might we address unintentional negative effects of AI explanations (e.g., inadvertently triggering cognitive biases that lead to over-trust)?
- What steps should we take to hold organizations/creators of XAI systems accountable and prevent "ethics washing" (the practice of ethical window dressing where 'lip service' is provided around AI ethics)?
- From an AI governance perspective, how can we address perverse incentives in organizations that might lead to harmful effects (e.g., privileging growth and AI adoption above all else)?
- How do we address power dynamics in the XAI ecosystem to promote equity and diversity?
- What are issues in the Global South that impact Human-centered XAI? Why? How might we address them?

4 WORKSHOP LOGISTICS

Pre-Workshop plans: Our pre-workshop plans serve three goals: **advertising** (to raise awareness and receive strong submissions), **building community**, and **recruiting expert reviewers.** We will build on effective strategies that have a proven track-record for past workshops we have organized. ***First,*** for advertising, we will use an integrated advertisement strategy that has two components– social media and mailing lists. The organizing committee has shared membership across many relevant disciplines like HCI, AI, NLP, Ubiquitous Computing, Sociology, Anthropology, Psychology, and Public Policy. We will primarily use Twitter (altogether the committee has more than 60,000 Twitter followers) and LinkedIn to advertise the workshop and engage with prospective participants. Beyond social media, we will distribute the Call for Papers through mailing lists (e.g., CHI, IUI, NeurIPs, AAAI, etc.). ***Second,*** in terms of community building, we will lean on two avenues– our existing online community on Discord as well as social media. We are fortunate to have a thriving and diverse online community on Discord, which started and continued from the first HCXAI workshop. We will encourage community-driven activities from ex-participants to engage with prospective participants. In addition to Discord, we plan to utilize the engagement through the advertisements on social media to expand our community. ***Third,*** we plan to recruit a Program Committee (PC) with expert reviewers similar to what we did for the first workshop. Based on past data, we expect more than 50 submissions, which entail having a core group of reviewers beyond that in the organizing committee will be instrumentalIn all of our efforts, we will prioritize diversity of perspectives and

representation in an effort to make the workshop as accessible and equitable as possible.

Workshop Mode: fully virtual. To promote equitable participation, we will host a fully virtual workshop. We have reached this decision based on consultation with different CHI stakeholders around Covid-related complexities like health & safety, global vaccine inequities, US-border closures, and visa restrictions. Moreover, we are persuaded by the advocacy of AccessSIGCHI [2] that promotes fully virtual events to lessen inequities arising from a hybrid format. With a virtual workshop, we can also broaden participation since travel costs and visas become less relevant. In 2021, the virtual format allowed us to *broaden participation* globally resulting in strong attendance from participants in the Global South. Moreover, the virtual mode allows us to continue the best practices derived from the 2021 edition. As we did it in the past, we will work with every participant to ensure inequities around internet and technological access is mitigated as much as possible, from allowing prerecorded presentations to archiving sessions for *asynchronous engagement*. This year, we expect around **100-125 participants**. Given last year's performance (with >100 participants), we are confident of facilitating in-depth discussions at this scale. We have also made operational adjustments from lessons learned last year, which can only improve the participant experience.

Website, Discord Server, and Asynchronous Engagement: Our website[1] provides a rich source of information and engagement for the workshop, from keynotes to expert panel discussions, from paper presentations to downloadable proceedings. Given the archival nature of the website, it has served as a *key portal for increased community engagement beyond the workshop* including new members who are likely to be future workshop participants. At the time of proposal submission, this website hosts content from 2021, which will be updated for 2022 (if accepted) while maintain access to past materials. Beyond the website, we have setup a Discord Server that serves as an online space for discussions before, during, and after the workshop. Given the virtual nature of the proposed workshop, Discord will host our participants virtually. As we outline below (in Workshop Plans), we will use a combination of Zoom and Discord for the workshop. To foster effective management, we have devoted significant resources to configure the Discord server in a way that allows access-based control with different roles like workshop participants, organizers, keynote speakers and panelists, etc. We also have a (virtual) registration desk to ensure registered participants get access to workshop related activities. This registration desk combined with the access-based controls solve a key problem for virtual workshops around assigning proper permissions. Taken together, **the website and the Discord server, affords effective asynchronous engagement**. In the past, participants have used Discord to engage asynchronously on discussions or catch up on missed presentations using the website. Beyond asynchronous avenues, we will use Zoom for live presentations. In the past, participants appreciated its transcription features for **increased accessibility**.

[1] https://hcxai.jimdosite.com/

5 WORKSHOP STRUCTURE

The workshop will take place through **two 4-hour online sessions (including breaks) on two subsequent days** (Table 1 outlines the key activities). Tentatively, the sessions will run from 1300-1700 ET, which was previously preferred by participants to accommodate different time zones. Once we finalize proceedings, we will collectively decide a final time with our participants. Approximately *two weeks before the start of the workshop,* we will share reading materials (e.g., past proceedings and recent impactful HCXAI papers). We will also ramp up social media engagement reusing our #HCXAI hashtag. Through a dedicated Discord channel, participants will have a chance to introduce themselves and begin engaging with each other. Since online events struggle from instantaneous rapport building, prolonging the introductory phases has shown to be effective in promoting conversations. If previous submission volume holds, we will have *two tracks*– full presentations (about 33%) and rapid-fire poster presentations (about 67%). Presentations happen in Zoom while the discussion happens on dedicated channels in Discord. This combination not only promoted a smooth experience (without cluttering the chat on Zoom calls) but also allowed for asynchronous engagement. Moreover, speakers appreciated being able to continue the conversation threads in Discord even after their talks are over.

On **Day 1**, we will begin with a *brief introductory session* that aligns participants with the workshop goals, outlines key activities, and introduces the organizers. Next, we will have a *keynote from an invited speaker* who is a thought leader at the intersection of AI and HCI. Last year, Tim Miller was our keynote speaker. The rest of day will include sessions for*full presentations and poster presentations*. We will have breaks between sessions to reduce video call fatigue. To *wrap up Day 1*, we will have a *virtual 'Happy Hour'* where participants can choose to stay, network, and engage in fun activities (like at-home scavenger hunts). We plan to host these activities through Discord and integrated platforms like WonderMe. Before the workshop, we will evaluate any new platforms to ensure *accessibility* is maintained.

Day 2 mainly involves panel discussions and group activities. It kicks off with an *expert panel discussion* with invited speakers from diverse disciplines that contribute to XAI. Next, there could be a *short presentation session* to accommodate the remaining presentations. Then, *group discussion* takes place. Discussion topics will be crowd-sourced and curated by the organizing committee. In the past, we have gathered them through surveys prior to the workshop. These topics can be shared before the workshop to allow for group formation *before* we go live. This way, we can minimize coordination challenges, which has helped us keep the workshop on time. Breakout rooms on Discord and/or Zoom will be provided (max 6 people per room). In terms of the activities, it will be a combination of community-driven and organizer-curated list. We can do *"papers from the future"*, a popular activity where participants engage in design fiction and envision future issues or opportunities with XAI. After deliberation, they generate a paper topic and abstract. This activity has already lead to many "offline" collaborations and real papers from the participants. Another option includes *generating actionable experimental designs or frameworks* (using Miro boards or shared Google slides). We hope this activity will also

Table 1: Tentative workshop structure, suggesting two 4-hour sessions (including breaks) on two subsequent days, as well as asynchronous activities before and after the workshop.

Start	End	Duration	Session
Before the Workshop			
-	-	2 weeks	Participants introduce themselves in the Discord channel and have access to provided workshop-related materials
*Workshop **Day 1***: 1300–1700ET			
13:00	13:30	30min	Introduction of workshop organizers, (remaining) participants, topics, and goals
13:30	14:30	60min	Keynote by invited speaker, including discussions
10 min break			
14:40	15:30	50min	Video presentations of position papers
15:30	17:00	90 min	Position paper poster session and networking
*Workshop **Day 2***: 1300–1700ET			
13:00	14:30	90min	Panel presentations and panel discussion
10 min break			
14:40	16:10	90min	Breakout group work
10 min break			
16:20	16:45	25min	Break-out group findings presentations
16:45	17:00	15min	Closing ceremony & Wrap-Up
After the workshop			
-	-	-	Results summary posted on workshop website & initiating follow-up activities

foster collaboration beyond the workshop. *After the group activity*, the participants regroup to share their discussions with quick "3-minute lightning talks". In the *closing ceremony*, we wrap up the workshop with a short presentation summarizing the work from the two days and acknowledge *impactful* position papers submitted. We also highlight areas of future work and propose ways to keep engaged with the HCXAI community through Discord and beyond.

Post Workshop Plans. We have a five-part plan. First, to continue community building, we plan to continue the conversation on Discord as we have done in the past. Second, our 2021 CHI workshop has already led to an accepted Special Issue (SI) in ACM Transactions on Intelligent Systems. The timeline of the SI allows us to invite strong submissions from the workshop to expand and submit to the journal. Third, we plan to use the website as an archival repository of workshop contents, which will hopefully continue to foster conversations and recruit new community members. Fourth, we will invite participants to write-up *synthesis papers* that could be published at ACM Interactions or Communications of the ACM and focused on open research areas and grand challenges in HCXAI. Last, if there is a critical mass of interested participants, we will explore transforming the workshop to a new conference in the future (similar to how FAT* workshops lead to ACM FAccT conference).

6 ORGANIZERS

The Organizing Committee is uniquely positioned to execute the visions of the workshop. We are a global team spanning industry and academia and bridging relevant XAI threads like AI, HCI, Sociology, Public Policy, and Psychology. Beyond hosting a previous version of this workshop, we have extensive organizational experience in HCI and AI venues.

Upol Ehsan is a doctoral candidate in the School of Interactive Computing at Georgia Tech. Existing at the intersection of AI and HCI, his work focuses on explainability of AI systems, especially for non-AI experts, and emerging AI Ethics issues in the Global South. He is also an affiliate at the Data & Society Research Institute. His work received multiple awards at ACM CHI and HCII. His work has pioneered the notion of Rationale Generation in XAI and also charted the vision for Human-centred XAI. Along with serving in multiple program committees in HCI and AI conferences (e.g., DIS, IUI, NeurIPS), he was the lead organiser for the *first* CHI workshop on Human-centred XAI..

Philipp Wintersberger is a researcher at the research center CARISSMA/THI. He obtained his doctorate in Engineering Science from Johannes Kepler University Linz specializing Human-Computer Interaction and Human-Machine Cooperation. He worked 10 years as a software engineer/architect before joining the Human-Computer Interaction Group at CARISSMA/THI to research in the area of Human Factors and Driving Ergonomics. His publications focus on trust in automation, attentive user interfaces, transparency of driving algorithms, as well as UX/acceptance of automated vehicles and have received several awards in the past years.

Q. Vera Liao is a Principal Researcher at Microsoft Research Montreal. Her current interest is in human-AI interaction and explainable AI, with a focus on bridging state-of-the-art AI technologies and user-centered design practices. She serves as the Co-Editor-in-Chief for Springer HCI Series, on the Editorial Board of International Journal of Human-Computer Studies (IJHCS) and ACM Transactions on Interactive Intelligent Systems (TiiS), and has been on the Organizing Committee for IUI 2019 and CSCW 2021. She actively organizes events that connect the HCI and AI communities, including several workshops and a panel at CHI, IUI and CSCW.

Elizabeth Anne Watkins is a Postdoctoral Fellow at Princeton University, with dual appointments at the Center for Information Technology Policy and the Human-Computer Interaction group,

and is also an affiliate with the AI on the Ground group at the Data and Society Research Institute. Trained as an organizational sociologist, her focus is on the oft-invisible articulation labor performed by humans to sustain systems of algorithmic decision-making. She has a special interest in the sociotechnical nexus of work, privacy, risk, and security. She's published or presented her research at CSCW, FAccT, and AIES, organized two workshops at CHI, and recently won Best Paper at the workshop on Transparency and Explanation in Smart Systems (TExSS) at IUI.

Carina Manger is a researcher at the research center CARISSMA/THI. Before she joined the Human-Computer Interaction Group, she obtained degrees in Psychology and Human Factors Engineering and worked on intelligent user interfaces in the automotive industry. Her current research concerns experimental user studies in simulated environments, with a strong focus on AI Explanations in automated driving. Her research approach aims to identify the underlying mental model of the user and is driven by theories from cognitive science and psychology.

Hal Daumé III is is a Perotto Professor in Computer Science and Language Science at the University of Maryland, College Park; he has a joint appointment as a Senior Principal Researcher at Microsoft Research, New York City. His primary research interest is in developing new learning algorithms for prototypical problems that arise in the context of natural language processing and artificial intelligence, with a focus on interactive learning and understanding and minimizing social harms that can be caused or exacerbated by computational systems. He has been program co-chair for ICML 2020 and for NAACL 2013. He was an inaugural diversity and inclusion co-chair at NeurIPS 2018.

Andreas Riener is professor for Human-Machine Interaction and Virtual Reality at Technische Hochschule Ingolstadt (THI) with co-appointment at the CARISSMA Institute of Automated Driving. He is program manger for User Experience Design and leads the UX/usability research and driving simulator labs. In 2017, he founded the interdisciplinary Human-Computer Interaction Group. His research interests include HF/ergonomics, adaptive UIs, driver state assessment, and trust/acceptance/ethics in automated driving. Andreas is steering committee co-chair of ACM AutomotiveUI and chair of the German ACM SIGCHI chapter.

Mark Riedl is a Professor in Georgia Tech's College of Computing and Associate Director of the Machine Learning Center at Georgia Tech. His research focuses on making agents better at understanding humans and communicating with humans. His research includes commonsense reasoning, story telling and understanding, explainable AI, and safe AI systems. He is a recipient of an NSF CAREER Award and a DARPA Young Faculty Award.

7 CALL FOR PARTICIPATION

AI-powered decisions increasingly pervade consequential domains of our lives in high-stakes domains (healthcare, finance, legal). Explainability has been sought as primary means, even fundamental rights, for people to understand, contest to foster equitable and just Human-AI collaboration. Although explainable AI (XAI) has been a fast-growing field, there is no agreed-up definition of, let alone methods to evaluate and guidelines to create XAI technologies. Discussions to chart the domain and shape these important topics call for human-centered and sociotechnical perspectives. In this workshop, we offer a junction point of cross-disciplinary stakeholders of the XAI landscape– from designers to engineers, from researchers to end-users. The goal is to examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological and technical levels. Consequently, we call for papers up to 6 pages excluding references that address topics involving the who (e.g., relevant diverse stakeholders), why (e.g., social/individual factors influencing explainability goals), or where (e.g., diverse application areas or evaluation). Papers should follow the CHI Extended Abstract format and be submitted through the workshop's submission site [2]. All accepted papers will be presented, provided at least one author attends the workshop and registers at least one day of the conference. Further, contributing authors are invited to provide their views in form of short panel discussions with the workshop audience. With an effort towards an equitable discourse, we particular welcome participation from the Global South and from stakeholders whose voices are under-represented in the dominant XAI discourse.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation under Grant No. 1928586 and by the FH-Impuls program of the German Federal Ministry of Education and Research (BMBF) under Grant Number 13FH7I06IA (MIRASOFT).

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] AccessSIGCHI. 2021. AccessSIGCHI statement on Chi 2022: Virtual or hybrid. https://docs.google.com/document/d/1mMmCXf1HT_7SNlcNpSS7U466WfVzuvRueh8HSSOOUWk/edit

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Wiebe E Bijker, Thomas P Hughes, Trevor Pinch, et al. 1987. The social construction of technological systems.

[5] 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*. 275–285.

[6] 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*. 1–19.

[7] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).

[8] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *arXiv preprint arXiv:2002.01092* (2020).

[9] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[10] 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*. 1–6.

[2] https://hcxai.jimdosite.com/

[11] Madeleine Clare Elish and Elizabeth Anne Watkins. 2020. Repairing Innovation: A Study of Integrating AI in Clinical Care. Data & Society.

[12] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2020).

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[14] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[15] Brittany E Holthausen, Philipp Wintersberger, Bruce N Walker, and Andreas Riener. 2020. Situational trust scale for automated driving (sts-ad): Development and initial validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 40–47.

[16] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

[17] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[18] Margot E Kaminski. 2019. The right to explanation, explained. *Berkeley Tech. LJ* 34 (2019), 189.

[19] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[20] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[21] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[22] Donald MacKenzie. 2018. Material Signals: A Historical Sociology of High-Frequency Trading. *Amer. J. Sociology* 123, 6 (2018), 1635–1683. https://doi.org/10.1086/697318

[23] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[24] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv* (2018), arXiv–1811.

[25] Johanna D Moore and William R Swartout. 1988. *Explanation in expert systemss: A survey.* Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

[26] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review* 2, 1 (31 3 2020). https://doi.org/10.1162/99608f92.6ed64b30 https://hdsr.mitpress.mit.edu/pub/7z10o269.

[27] Edward Shortliffe. 2012. *Computer-based medical consultations: MYCIN*. Vol. 2. Elsevier.

[28] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[29] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.