Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs)

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

Elizabeth Anne Watkins
Intel Labs, Intelligent Systems Research
CA, USA
elizabeth.watkins@intel.com

Philipp Wintersberger
Univ. of Applied Sciences Upper Austria / TU Wien
Austria
philippwintersberger@gmail.com

Carina Manger
Technische Hochschule Ingolstadt (THI)
Ingolstadt, Bavaria, Germany
carina.manger@thi.de

Sunnie S. Y. Kim
Princeton University
Princeton, New Jersey, USA
sunniesuhyoung@princeton.edu

Niels van Berkel
Aalborg University
Denmark
nielsvanberkel@cs.aau.dk

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

Human-centered XAI (HCXAI) advocates that algorithmic transparency alone is not sufficient for making AI explainable. Explainability of AI is *more* than just "opening" the black box — *who* opens it matters just as much, if not more, as the ways of opening it. In the era of Large Language Models (LLMs), is "opening the black box" still a realistic goal for XAI? *In this **fourth** CHI workshop on Human-centered XAI (HCXAI)*, we build on the maturation through the previous three installments to craft the coming-of-age story of HCXAI in the era of Large Language Models (LLMs). We aim towards actionable interventions that recognize both affordances and pitfalls of XAI. The goal of the fourth installment is to question how XAI assumptions fare in the era of LLMs and examine how human-centered perspectives can be operationalized at the conceptual, methodological, and technical levels. Encouraging holistic (historical, sociological, and technical) approaches, we emphasize "operationalizing." We seek actionable analysis frameworks, concrete design guidelines, transferable evaluation methods, and principles for accountability.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3636311

ACM Reference Format:
Upol Ehsan, Elizabeth Anne Watkins, Philipp Wintersberger, Carina Manger, Sunnie S. Y. Kim, Niels van Berkel, Andreas Riener, and Mark O. Riedl. 2024. Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3636311

1 INTRODUCTION

When the initial vision of Explainable AI (XAI)—a field concerned with developing techniques, concepts, and processes that can help stakeholders understand the reasons behind the AI system's decision-making [12, 19]—was articulated, a popular framing was to "open" the (proverbial) "black-box" of AI [6, 23], so that we could see inside of it, figure out what it was doing, why it was doing it, and if it was doing it correctly. With the proliferation of Large Language Models (LLMs), *is "opening the black box" a realistic expectation in XAI anymore?*

When we consider an LLM-powered service such as ChatGPT, GPT-4, Bing Chat, Bard, LaMDA, or PaLM, what prospects are there for "opening" the black-box of AI? These models have hundreds of billions of parameters, all acting in conjunction to generate a distribution of possible words to choose from to build a response, word by word. If we had access to all the weights, could we interpret and explain the model? If we had access to the parameters of a model and the activation values for an input could we interpret and explain the model? Many of these models run behind APIs that prevent such inspection, but even if we could access this information, the raw values of weights and activations are meaningless to most people without synthesizing some visualization or text summarization that provides a lay-understandable analysis of the internal operations of the system and how the results were generated by the system. Consider OpenAI's work on interpreting the patterns that cause

individual neurons to activate [5]. How would knowing what causes neuron #2142 to activate help people, especially non-AI experts, know how to better use ChatGPT? What actionable information from this neural activation pattern can we use meaningfully?

If opening the black-box is no longer a meaningful goal, is XAI doomed to fail? No, this is where *Human-Centered Explainable AI (HCXAI)* [11] can help. HCXAI encapsulates the philosophy that not everything that is important lies inside the black box of AI. Critical insights can lie outside it *because that's where the humans are.* HCXAI argues that XAI is *more* than just "opening" the black-box— *who* opens it matters just as much, if not more, as the ways of opening it [11]. In the era of LLMs, XAI approaches that center the human and the sociotechnical components of AI systems may prove even more resilient than approaches that focus only on technical attributes. Given that real-world AI systems exist in sociotechnical settings, it takes more than just algorithmic transparency to make them explainable [14]. Increasing the aperture of XAI can help us focus on the most important part: *who* the human(s) is (are), what they are trying to achieve in seeking an explanation, and how to design XAI techniques that meet those needs. Given that XAI is as much of HCI's problem as it is AI's, CHI continues to be the venue to address the human side of XAI.

In this ***fourth* installment** of the Human-centered Explainable AI (HCXAI) workshop at CHI, we build on the groundwork of the first three workshops in 2021, 2022, and 2023 [13–15] with a combined attendance of over 350 participants from over 18 countries. In 2021, the conversation was centered around the pressing need to systematically understand XAI evaluation methods. In 2022, the community increased in diversity and depth, and the conversation matured to address XAI harms and devise mitigation strategies. In 2023, there was a "turn to the infrastructure" with deep discussions around how we should focus on different stages of the AI's lifecycle and leverage "seamful information" [9] to augment XAI. An outcome of the community engagement was a Journal Issue on Human-Centered XAI (ACM Transactions on Interactive Intelligent Systems). In 2024, for the 4^{th} workshop, we continue the *coming of age* journey for HCXAI in the era of LLMs and beyond. We continue to serve as a junction point for cross-disciplinary stakeholders to tackle HCXAI 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 HCXAI community of diverse stakeholders.

2 HUMAN-CENTERED XAI: BACKGROUND AND OPPORTUNITIES

To harmonize the different threads of work in HCXAI, we adopt the analytic lens of "Social Construction of Technology" (SCOT) [4], a theory in Science & Technology Studies that centers human action as foundational in developing the shape and function of technology. We acknowledge that diverse *relevant social groups* (e.g., different stakeholders such as researchers, or policymakers) 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, 2, 22]. Some have defined explainability as an AI systems' decisions being *easy to understand* by people [2, 17], 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* [12] that communicate an opaque model's decisions in a way that is 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.

Our workshop emphasizes that the *who in XAI* matters due to the diverse needs of involved stakeholders including data scientists, decision-makers, regulatory agencies, and end-users, and brings this diversity to the forefront. In HCXAI, we not only ask about for *whom* an explanation is created, but also *why* [11], since explanations are requested for a wide range of purposes such as *trustworthiness, causality, fairness, accessibility, interactivity, or privacy-awareness* [2]. 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, policymakers, drivers, etc. require different forms of explanations. Ehsan et al. [10] highlighted how users with different backgrounds ascribe different meanings to the same form of explanations. Besides the *why, who,* and *"where"*, the application domain or context also play important roles. For example, recent work has introduced XAI features into model development tools [16], AI-assisted decision-support tools [24], and model fairness evaluation [8].

Given the progress so far, it is imperative to continue the critically constructive narrative around HCXAI to address intellectual blind spots and propose human-centered interventions. The goal is *not* to impose a normativity but to systematically articulate the different *interpretive flexibilities* of each *relevant social group* in XAI. This allows us to make actionable progress at all three levels–conceptual, methodological, and technical.

3 GOALS OF THE WORKSHOP & AREAS OF INTEREST

Broadly, the **goals** are to (1) extend the constructive dialogue around *operationalizing* HCXAI at the conceptual, methodological, and technical levels and (2) expand the HCXAI community. Operationalization can include actionable analysis frameworks, concrete design guidelines, transferable evaluation methods, and principles for accountability.

For the 2024 edition of the workshop, we aim to balance breadth and depth. In terms of foundational areas, we remain interested in topics like *sociotechnical aspects* of XAI, *human-centered evaluation* techniques, *responsible use* of XAI. Beyond these, we aim to expand on the following areas (inspired by discussions from the first three workshops) [13–15]: exploring how *power dynamics manifest in XAI* and how we can *foster accountability and avoid "ethics washing"* in XAI. **In the era of LLMs, all of these questions take on new attributes, new actors, and new urgency.** The open-source availability of some LLMs and ease of building "wrappers" for different task applications means AI inference and prediction are being pushed into ever more domains. Further, the anthropomorphic design of these tools and new capabilities means they are

beginning to engage with humans in areas of life previously considered too specialized, too creative, or too intimate for AI to tread. Now, more than ever, humans need good, accessible transparency into the systems with which they are interacting [20]. There are growing demands on the field of XAI to equip human beings with the understanding they need to use these tools [18], in ways that are safe, responsible, and productive [7]. HCXAI is more urgent than ever before. The following list of guiding questions is *not* an exhaustive one; rather, it is provided as a source of inspiration for position papers:

- From an HCXAI angle, how should we **think about the explainability of LLMs** given the challenges of translating multi-billion-parameter models into meaningful and accessible explanations for lay users?
- Just because LLMs can respond to why-questions, does that mean **LLMs can "explain" themselves**?
- All LLMs hallucinate. How might we use HCXAI to **detect hallucinations** & mitigate negative effects?
- How do we address the **power dynamics** in XAI? Whose "voices" are represented in AI explanations? Who gets to say what explanations users see?
- How should we **practice Responsible AI** when it comes to XAI? How might we mitigate risks with explanations, what risks would those be, and how does risk mitigation map to different stakeholders?
- How can we **create XAI Impact Assessments** (similar to Algorithmic Impact Assessments)?
- How should organizations/creators of XAI systems be **held accountable** to prevent "ethics washing" (the practice of ethical window dressing where "lip service" is provided around AI ethics)?
- How might we design XAI systems that are **dark-pattern resistant**—how might we hinder AI explanations from being weaponized for over-reliance or over-adoption?
- Can we reconcile the tension between XAI and privacy? If yes, how? If no, why?
- Given the contextual nature of explanations, what are the potential **pitfalls of standardized evaluation metrics**? How might we take into account the who, why, and where in the evaluation methods?
- How might **explanations be designed for actionability**, to provide action-oriented nudges to enable users to become better collaborators with AI systems?
- How might we address **XAI issues in the Global South (Majority World)**?
- How should we think about **explanations in physical systems** (e.g., self-driving cars) vs. those in non-physical ones (e.g., automated lending)? Are there effectively the same? Are they different?

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 speakers & expert reviewers.** To achieve these goals, we will use effective strategies that have a proven track record from previous events. ***First,*** for advertising, we will use an integrated advertising strategy that has two components - social media and mailing lists. The organizing committee has shared membership across many relevant disciplines like HCI, AI, NLP, Sociology, Psychology, and Public Policy. We will primarily use Twitter and LinkedIn to advertise leveraging our network of over 60,000 combined followers. Beyond social media, we will distribute the Call for Papers through mailing lists (e.g., CHI, IUI, NeurIPs, AAAI). ***Second,*** for community building, we will use two avenues - our existing online community on Discord and social media. We are fortunate to have a thriving online community on Discord, which started and continued from the first HCXAI workshop, and which we will continuously foster to raise engagement. We will encourage community-driven activities from ex-participants to engage with prospective participants. In addition to Discord, we plan to utilize participation through social media advertisements. ***Third,*** as in previous years, we will recruit a Program Committee (PC) to handle at least 50-60 submissions (based on prior data) and recruit thought leaders as keynote speakers. So far, *we have a successful track record of recruiting thought leaders from both inside and outside XAI.* In 2021, we had Tim Miller, a key voice in XAI. In 2022, Tania Lombrozo, whose work in the psychology of explanations is seminal to XAI, joined us. In 2023, we had a lively debate from different disciplines– Andres Paez (philosopher), Tania Lombrozo (psychologist), and Tim Miller (computer scientist). For 2024, we already have a short list of speakers from diverse threads (e.g., AI ethics, policy) and are confident of successfully recruiting them. In all 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: Hybrid Event. HCXAI 2024 will be held as a hybrid event. The committee has a combined experience of organizing over 30 workshops and conferences (virtual, in-person, and hybrid [3]). We will use our experience to avoid common pitfalls and leverage the strengths of both worlds . We have introduced parity and equity in virtual participation in previous hybrid workshops we have organized by taking a *virtual-first* design approach (as opposed to the usual *in-person first*). For instance, everyone logs into Zoom regardless of whether they are in-person or not. This way the virtual participants do not feel alienated and can enjoy virtual co-presence with everyone. When people present, instead of pointing a camera at the stage where screen legibility suffers, participants share their screens on Zoom and present. Given everyone is on Zoom, there are no A/V issues in the room. To make up for time differences, we also upload materials ahead of time so that participants can watch them asynchronously and submit questions ahead of time. These strategies will allow us to streamline a comfortable experience for all participants. Ultimately, the time zone (time difference: 6h to New York, 12h to Berlin) of the conference demands a carefully designed approach. The strategy will allow us to *broaden participation* globally and from participants in the Global South due to travel costs and visa issues. As in previous years, we expect around **100-125 participants** (40-50 in-person, the rest virtual). Given our track record, we are confident in facilitating in-depth discussions at this scale. *In the event CHI moves online, given our past experience, we are also fully capable of hosting the workshop online.*

Website, Publishing Workshop Proceedings, Discord Server, and Asynchronous Engagement: Our **website** (https://hcxai.jimdosite.com/) provides a rich source of information and engagement for the workshop, from keynotes to hosting proceedings. Given the archival nature of the website, it has served as a *key portal for increased community engagement beyond the workshop*. At the time of proposal submission, this website hosts content from 2021 to 2023, which will be updated for 2024 (provided acceptance). Beyond the website, we have set up a **Discord server** that serves as an online space for discussions before, during, and after the workshop. Online participants will be able to reach the provided activities via Discord. Taken together, **the website and the Discord server, affords effective asynchronous engagement**. In the past, participants have used Discord to engage asynchronously in discussions or catch up on missed presentations using the website. Beyond asynchronous avenues, we will use Zoom (including live transcription) for live streaming of the in-person event at CHI'24.

5 WORKSHOP STRUCTURE

The workshop is planned as a full day event consisting of **two 4-hour sessions (including coffee breaks)** (Table 1 outlines the key activities). Tentatively, the sessions will run from 9am-1pm and 2pm-6pm local time in Hawaii. To account for time differences and best integrate remote participants, we will perform group activities directly in the morning (allowing participants calling in from Europe and the US mainland to join) and in the late afternoon (to include participants from Asia). For these activities, we will combine relaxed group discussions with networking possibilities. Directly before and after lunchtime, keynotes, paper, and poster presentations will be held on-site. Remote participants can watch pre-recorded paper and poster presentations during the day before the event. The keynote as well as the Q&A sessions will be recorded and immediately posted on the Discord server. We briefly outline the course of events for four types of attendees (on-site, remote from Europe, Asia, and the US) which will allow participants from all over the globe to engage equitably:

- **On-site participants** will use the first session to introduce themselves and discuss/network with both other on-site and remote participants. Then, they experience the keynote and the first two paper sessions. After lunch, the third paper session and the poster presentations will be hosted. The day concludes with another set of group discussions and networking activities at the end.
- **European participants** can join the discussion/networking activities and the keynote in their evening. Since they had the chance to watch the pre-recorded presentations and post questions already during the day, they can enjoy a relaxed night of sleep. In the morning, they can again participate in the concluding discussions before watching the just-before posted recordings of the paper Q&As.
- **Asian participants** may get up early to join the paper presentations and poster sessions online. Shortly before lunch, they can join the concluding networking and discussion rounds.
- **US participants** can join the entire event online during their afternoon and late evening. Still, they can take advantage of all online materials just as participants from other remote locations.

Once we finalize the proceedings, we will collectively decide a final time with our participants. *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 using our #HCXAI hashtag. Through a dedicated Discord channel, participants will have a chance to introduce themselves and begin engaging with each other. Since online and hybrid events struggle with instantaneous rapport building, prolonging the introductory phases has shown to be effective in promoting conversations. If the previous submission volume holds, we will have *three tracks*– full presentations (about 33%), rapid-fire poster presentations (about 67%), and the creative short-form videos-only track ("The Sanities & Insanities of XAI: Provocations & Evocations"). Presentations happen in Zoom while the discussion happens on dedicated channels in Discord. This combination promoted a smooth experience (without cluttering the chat on Zoom calls) and asynchronous engagement—speakers appreciated being able to continue the conversation on Discord even after their talks are over.

In **Session 1**, we will begin with a *brief introductory session* that aligns participants with the workshop goals, outlines key activities, and introduces the organizers. This is followed by a first networking and discussion section in form of a *'Happy Hour'* where both on-site and online participants can network and engage in fun activities (like at-home scavenger hunts). Next, we will have a *fireside chat (interactive keynote)* with a thought leader at the intersection of AI and HCI (such as Tim Miller and Tanja Lombrozo in previous years). Instead of a keynote monologue, past experience suggests that an interactive fireside chat format (15-min presentation with 45-min Q&A) simultaneously facilitates engagement from the audience while reducing the speaker's burden of preparing a long presentation. The rest of the session will include presentations of papers and posters. We will have breaks between sessions to reduce fatigue. We plan to host these activities through Discord and integrated platforms like WonderMe. All platforms will be evaluated for *accessibility* before adoption.

Session 2 involves panel discussions and group activities. It starts with an *expert panel discussion* with invited speakers from diverse disciplines that contribute to XAI. Potentially, there could be a *short presentation session* to accommodate the remaining presentations. Then, *group discussion* takes place. Discussion topics will be crowd-sourced (via surveys prior to the workshop) and curated by the organizing committee. Teams are split into breakout rooms on Discord or sit together on-site (max 6 people per room). We have done a range of group activities in the past - from groups brainstorming *"papers from the future"* with thought-provoking ideas to *"news headlines in the near and far future"* where participants engage in design fiction and envision future issues or opportunities with XAI. These activities have led to many collaborations and papers from the participants. *After the group activity*, the participants regroup to share their discussions with quick "2-minute lightning talks". In the *closing ceremony*, we wrap up the workshop with a short presentation summarizing the work and recognize *impactful* position papers submitted. We also highlight areas of future work

Table 1: Tentative workshop structure, suggesting two 4-hour sessions (including breaks), 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 **Session 1***: 09:00–13:00 Hawaii Local Time			
09:00	09:30	30min	Introduction of workshop organizers, topics, and goals
09:30	10:30	60min	Networking and Participant Introductions
30 min break			
11:00	12:00	60min	Keynote by invited speaker, including discussions
12:00	12:30	30min	Position paper session
12:30	13:00	30min	Poster session
1 hour break			
*Workshop **Session 2***: 14:00–18:00 Hawaii Local Time			
14:00	15:30	90min	Panel presentations and panel discussion
10 min break			
15:40	17:10	90min	Break-out group work
10 min break			
17:20	17:45	25min	Break-out group findings presentations
17:45	18:00	15min	Closing ceremony & Wrap-Up
After the workshop			
-	-	-	Results summary posted on workshop website & initiating follow-up activities

and propose ways to keep engaged with the HCXAI community through Discord and beyond.

Post Workshop Plans. We have a four-part plan. First, to continue community building, we plan to continue the conversation on Discord as we have done in the past. Second, we plan to use the website as an archival repository of workshop content, which will hopefully continue to foster conversations and recruit new community members. Third, 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 the 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 previous versions 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. His work focuses on explainability of AI systems, especially for non-AI experts, and emerging AI Ethics issues in the Global South. His work received multiple awards at ACM CHI and HCII. His work has 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 organizer for the 2021, 2022, and 2023 CHI workshops on Human-centred XAI.

Philipp Wintersberger is a Professor of Interactive Systems at the University of Applied Sciences Upper Austria (Campus Hagenberg). His research addresses human-machine cooperation in safety-critical AI-driven systems. Currrently, he leads a group of researchers and Ph.D. students working on human-AI cooperation in multiple FWF and FFG-funded projects

Elizabeth Anne Watkins is a Research Scientist in the Social Science of AI at Intel Labs Intelligent Systems Research, where she serves on the Responsible AI Advisory Council. She was a Postdoctoral Fellow at Princeton University, with dual appointments at the Center for Information Technology Policy and the Human-Computer Interaction group. She has published or presented her research at CSCW, CHI, FAccT, USENIX, and AIES, and co-organized multiple CHI workshops including the 2022 and 2023 HCXAI workshops.

Carina Manger is a researcher at the research center CARISSMA/THI. Before joining 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.

Sunnie S. Y. Kim is a computer science PhD candidate at Princeton University, supported by the NSF Graduate Research Fellowship. She works on AI transparency and explainability to help people better understand and interact with AI systems. Her research is at the intersection of AI and HCI and has been published in both fields (e.g., CVPR, ECCV, CHI, FAccT). She also led the organization of the Explainable AI for Computer Vision Workshop (XAI4CV) at CVPR 2023.

Niels van Berkel is an Associate Professor at Aalborg University's Human-Centred Computing group. His research focuses on Social Computing and Human-AI interaction, with a special interest in decision-support. In addition to real-world evaluations, he has a keen interest in methodological contributions.

Andreas Riener is a professor for Human-Machine Interaction and VR at Technische Hochschule Ingolstadt with co-affiliation at the CARISSMA Institute of Automated Driving. His research interests include HF/ergonomics, adaptive UIs, driver state assessment, and trust/acceptance/ethics in mobility applications. Andreas is on the SC of ACM AutomotiveUI, chair of the German ACM SIGCHI chapter, and SC chair of "Mensch und Computer" (the German CHI).

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

Explainability is an essential pillar of Responsible AI. Explanations can improve real-world efficacy, provide harm mitigation levers, and serve as a primary means to ensure humans' right to understand and contest decisions made about them by AI systems. In ensuring this right, XAI can foster equitable, efficient, and resilient Human-AI collaboration. In this workshop, we serve as 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 position papers making justifiable arguments (up to 4 pages excluding references) that address topics involving the who (e.g., relevant diverse stakeholders), why (e.g., social/individual factors influencing explainability goals), when (e.g., when to trust the AI's explanations vs. not) or where (e.g., diverse application areas, XAI for actionability or human-AI collaboration, or XAI evaluation). Papers should follow the CHI Extended Abstract format and be submitted through the workshop's submission site (https://hcxai.jimdosite.com/). 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 the form of short panel discussions with the workshop audience. With an effort towards an equitable discourse, we particularly welcome participation from the Global South and from stakeholders whose voices are underrepresented in the dominant XAI discourse.

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] 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.

[3] Zoe M. Becerra, Nadia Fereydooni, Andrew L. Kun, Angus McKerral, Andreas Riener, Clemens Schartmüller, Bruce N. Walker, and Philipp Wintersberger. 2020. Interactive Workshops in a Pandemic…The Real Benefits of Virtual Spaces. *submitted to IEEE Pervasive Computing* (2020).

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

[5] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[6] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nature News* 538, 7623 (2016), 20.

[7] Teresa Datta and John P Dickerson. 2023. Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. *arXiv preprint arXiv:2303.06223* (2023).

[8] 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.

[9] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daume III. 2022. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *arXiv preprint arXiv:2211.06753* (2022).

[10] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I.-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv:2107.13509 [cs]* (July 2021). arXiv:2107.13509 [cs]

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

[12] 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.

[13] 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.

[14] 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 Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[15] Upol Ehsan, Philipp Wintersberger, Elizabeth A Watkins, Carina Manger, Gonzalo Ramos, Justin D Weisz, Hal Daumé Iii, Andreas Riener, and Mark O Riedl. 2023. Human-Centered Explainable AI (HCXAI): Coming of Age. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[16] 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).

[17] 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.

[18] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[19] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[20] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).

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

[22] 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.

[23] George Nott. 2017. Explainable artificial intelligence: Cracking open the black box of AI. *Computer world* 4 (2017).

[24] 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.