Human-Centered Explainable AI (HCXAI): Coming of Age

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

Philipp Wintersberger
Univ. of Applied Sciences Upper Austria / TU Wien
Austria
philipp.wintersberger@carissma.eu

Elizabeth Anne Watkins
Intel Labs, Intelligent Systems Research
CA, USA

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

Gonzalo Ramos
Microsoft Research
Kirkland, Washington, USA
goramos@microsoft.com

Justin D. Weisz
IBM Research
Yorktown Heights, New York, USA
jweisz@ibm.com

Hal Daumé III
University of Maryland & Microsoft Research
Oxford, England, USA

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 is an essential pillar of Responsible AI that calls for equitable and ethical Human-AI interaction. Explanations are essential to hold AI systems and their producers accountable, and can serve as a means to ensure humans' right to understand and contest AI decisions. Human-centered XAI (HCXAI) argues that there is more to making AI explainable than algorithmic transparency. 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 this third CHI workshop on Human-centered XAI (HCXAI)*, we build on the maturation through the first two installments to craft the coming-of-age story of HCXAI, which embodies a deeper discourse around operationalizing human-centered perspectives in XAI. We aim towards actionable interventions that recognize both affordances and potential pitfalls of XAI. The goal of the third 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 emphasize "operationalizing." Within our research agenda for XAI, 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**; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence.

ACM Reference Format:
Upol Ehsan, Philipp Wintersberger, Elizabeth Anne 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 (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3573832

1 INTRODUCTION

Explainability of AI systems is crucial for their accountability, which is crucial given their proliferation in high-stakes domains like healthcare [16], finance [20], and criminal justice [23]. Explainable AI (XAI) aims to provide human-understandable information for the system's behavior and processes [4, 11]. Despite the growth spurt in algorithmic approaches to "open" the black-box of AI, investigations on how people actually interact with AI explanations have found popular XAI techniques to be ineffective [3, 22, 25] and potentially risky [17, 24].

Human-centered XAI (HCXAI) argues that 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 [9]. Introducing the concept of Human-centered XAI, Ehsan and Riedl [10] observe that given that real-world AI systems exist in sociotechnical settings, it takes more than just algorithmic transparency to make them explainable [13]. Thus, explaining what is happening "inside the black box" often requires us to also understand things "outside the black box" [6, 18], requiring us to consider the entire AI lifecycle (vs. just the algorithm). To foster Responsible AI, we need to go beyond opening the black-box and towards Human-centered XAI. 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 ***third* installment** of the Human-centered Explainable AI (HCXAI) workshop at CHI, we build on the groundwork of the first two workshops in 2021 and 2022 [12, 13] with a combined attendance of over 220 participants from over 18 countries with over 110 submissions. In 2021, the conversation was centered around the

CHI EA '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573832

pressing need to go beyond algorithmic approaches and consider user-centered techniques in XAI. The workshop also served as a platform to start forming a community of diverse researchers, practitioners, and policymakers. In 2022, the community increased in diversity and depth, and the conversation matured to address more insidious, fine-grained, and nuanced aspects of XAI only visible with a human-centered lens. For instance, one paper session addressed how explanations in AI are *not* a golden bullet and how they could be weaponized to achieve dark patterns. An outcome of the community engagement was a Journal Issue on Human-Centered XAI (ACM Transactions on Interactive Intelligent Systems).

In 2023, for the *third* workshop, we use the metaphor of *coming of age* for HCXAI. In building out the "so what?" implications of the descriptive findings of prior workshops, this year we move towards action - not only regarding specific topics but also considering the future of the HCXAI community. Topic-wise, for example, now that we have identified the dark patterns of AI explanations, how can we design dark pattern-resistant XAI? Or, as we have shown user groups have distinct goals, how can design explanations to support them? Across issues, how can we take steps towards actionable interventions? Beyond addressing these specific questions, we will actively discuss the future of this growing community - for example, should it host its own conference in the future, and if so, how should such an event be organized so that the diverse needs of all members are properly supported?

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. Within our research agenda for XAI, we seek actionable analysis frameworks, concrete design guidelines, transferable evaluation methods, and principles for accountability.

2 HUMAN-CENTERED XAI: BACKGROUND AND OPPORTUNITIES

To harmonize the different threads of work in HCXAI, we adopt the analytic lenses of "Social Construction of Technology" (SCOT) [5], 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, 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, 4, 21]. Some have defined explainability as an AI systems' decisions being *easy to understand* by people [4, 15], and the term is often viewed more broadly than transparency or interpretable models [19]. This is illustrated by a growing area within the XAI community, which addresses *post-hoc explanations* [11] that communicate an opaque model's decisions in a way that is accessible for end users [19], 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 by the diverse needs of involved stakeholders including data scientists, decision-makers, regulatory agencies, and end-users to the forefront. This is in contrast to an algorithm-centered XAI approach that often privileges the algorithmic side. In HCXAI, we not only ask about for *whom* an explanation is created, but also *why* [9], since explanations are requested for a wide range of purposes such as *trustworthiness, causality, fairness, accessibility, interactivity, or privacy-awareness* [4]. 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. [8] highlighted how users with different backgrounds ascribe different meanings to the same form of explanations. Beside the *why, who*, and *"where"*, the application domain or context also plays important roles. For example, recent work has introduced XAI features into model development tools [14], AI-assisted decision-support tools [25], and model fairness evaluation [7].

The last two workshops have facilitated progress in all these areas in HCXAI. In 2021, we opened the stage towards HCXAI, which resulted in paper sessions addressing the *involvement of end users*, *explanation design*, and *theoretical frameworks*. The 2022 workshop took the conversation further and branched out into topics such as dark patterns and problems in XAI (*Concerns and Issues*), how end users perceive and directly engage with explanations (*Trust and Cooperation*), as well as how explanations can be tailored for individuals (*Human-centered explanation design*).

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 groups* 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

We aim to chart the HCXAI landscape from historical, sociological, and technological perspectives, and seek levers to make our findings actionable. Broadly, the **goals** are to (1) extend the critically constructive dialogue around *operationalizing* human-centered perspectives in XAI at the conceptual, methodological, and technical levels and (2) further expand, discuss, and draft the future of the HCXAI community. Operationalization can include aspects such as actionable analysis frameworks, concrete design guidelines, transferable evaluation methods, and principles for accountability. Bridging works from researchers, designers, and practitioners from the fields of XAI, HCI, psychology, machine learning, and social sciences, we want to craft the narrative of *coming of age* for HCXAI.

For the 2023 edition of the workshop, we want to balance breadth and depth. We want to build on the foundation while expanding our horizons. In terms of foundational areas, we are still interested in topics like *sociotechnical aspects* of XAI, *human-centered evaluation* techniques, *responsible use* of XAI. Beyond these, we want to expand on the following areas (inspired by discussions from the first two workshops) [12, 13]: exploring how *power dynamics manifest in XAI* (organizationally, communally, etc.), how we can *build resilient systems that resist dark patterns* of XAI, and how we can *foster accountability and avoid "ethics washing"* in XAI.

Beyond the usual papers track, this year we will add a creative way to join the workshop. Dubbed as *"The Sanities & Insanities of XAI: Provocations & Evocations"*, we will invite creative short-form (60-90 seconds) *video-only* submissions. The videos should speak to our areas of interest in a creative and provocative way. The aim is not to present academic work but to inspire discussions about the unknowns of XAI in an already unknown future. Participants can use science fiction, design fiction, speculative design, and other creative techniques to produce videos envisioning futures around XAI; for e.g., future public service announcements, poetry reading about dystopian or utopian future XAI scenarios, imagining what XAI would look like if the power dynamics in AI were different–what if AI were mostly developed in the Global South (instead of the Global North), what if we used indigenous ways of thinking about AI, how would HCXAI look like? Adopting a creative outlet can broaden participation from non-traditional groups such as artists, activists, and policymakers who have insightful perspectives but often face barriers to joining academic workshops.

The aforementioned areas can be situated from the foundational pillars of *who* (e.g., clarifying *who* the human is in XAI), *why* (e.g., how individual and social factors influence explainability goals), *when* (e.g., when to trust the AI's explanations vs. not) and *where* (e.g., explainability differences across application domains). The following list of guiding questions is *not* an exhaustive one; rather, it is provided as a source of inspiration for position papers:

- 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 address **value tensions** amongst different stakeholders in XAI?
- 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?
- How might we address **perverse organizational incentives** which could lead to harmful effects (e.g., privileging growth and AI adoption above all else)?
- Can we reconcile the tension between XAI and privacy? If yes, how? If no, why?
- How do **user characteristics** (e.g., profession, educational training) impact needs around explainability?
- When should we trust an AI's explanations and when should we ignore them? What are the characteristics that govern that **trust calibration**?
- 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?
- What are the **issues in the Global South (Majority World)** that impact Human-centered XAI? How might we address them?
- 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?
- **Explainability** is often considered at the software level. What should **hardware-based companies** that use AI need to consider about XAI?

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 the last two years. ***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 (the committee has more than 60,000 Twitter followers) and LinkedIn to advertise. 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. 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, Tim Miller, one of the most influential researchers in XAI joined us. In 2022, Tania Lombrozo, whose work in the psychology of explanations is seminal to XAI, joined us in an innovative fireside chat (instead of a keynote monologue). For 2023, we already have a short list of speakers from diverse threads of XAI (e.g., philosophy, 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: fully virtual. We plan to hold a fully virtual workshop for two major reasons. First, we promote equitable participation, and second, given the high number of workshop attendees, CHI workshop locations on-site could be too small. We have reached this decision based on consultation with different CHI stakeholders around COVID-related challenges, global vaccine inequities, and visa restrictions. Moreover, we are persuaded by the advocacy of AccessSIGCHI [2] that promotes fully virtual events to lessen inequities. In 2021 and 2022, the virtual format allowed us to *broaden participation* globally, resulting in strong attendance from participants in the Global South because travel costs and visas were non-issues. By going virtual, *we are not constrained by space restrictions prevalent in in-person settings.* As we have done in the past, we will work with each participant to ensure inequities around internet and technological access are mitigated as much as possible, from allowing prerecorded presentations to archiving sessions for *asynchronous engagement*. So far, there has been *zero cases* where participants could not attend fully due to technical reasons. Like past workshops, we expect around **100-125 participants**. Given our track record (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 will further improve the participant experience.

In 2023, for the third time in a row, CHI will conflict with Ramadan and Eid, observed by over 2 billion people globally, including many in our CHI community. In 2022, our virtual format allowed us to avoid the clash because we hosted the workshop after these dates. Since it was virtual, we did not need extra rooms, which entailed, there was no compulsion to host it during the CHI week. This freedom allowed us to be more inclusive and equitable, which boosted our participation. We will continue our tradition of putting the DEI principles of SIGCHI into practice.

Website, 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 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 and 2022, which will be updated for 2023 (provided acceptance) while maintaining access to past materials. Beyond the website, we have set up 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 that only participants who registered for CHI have access to workshop-related activities. This registration desk combined with the access-based controls solves 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 in 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**.

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). Data from previous years indicate that two days are optimal to adequately facilitate conversations without exhausting the participants. Tentatively, the sessions will run from 7pm-11pm CET/1pm-5pm ET, which was previously preferred by participants to accommodate different time zones. 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 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 **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 *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 day will include the newly created video session (*The Sanities & Insanities of XAI: Provocations & Evocations)* as well as *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 network and engage in fun activities (like at-home scavenger hunts). We plan to host these activities through Discord and integrated platforms like WonderMe. All platforms will be evaluated for *accessibility* before adoption.

Day 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

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, topics, and goals
13:30	14:30	60min	Keynote by invited speaker, including discussions
10 min break			
14:40	15:00	20min	Short-form creative videos: *The Sanities & Insanities of XAI: Provocations & Evocations*
15:00	16:30	90min	Position paper poster session and networking
16:30	17:00	30 min	Presentation of position papers
*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

Discord (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 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 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. 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 organizer for the 2021 and 2022 CHI workshops 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 in 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.

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, and was also an affiliate with the AI on the Ground group at the Data and Society Research Institute. She has published or presented her research at CSCW, FAccT, USENIX, and AIES, co-organized three workshops at CHI including the 2022 HCXAI workshop, and currently serves on the CHI subcommittee for Critical and Sustainable Computing.

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. Her research approach aims to identify the underlying mental model of the user and is driven by theories from cognitive science and psychology.

Gonzalo Ramos is a Principal Researcher working on Interactive Machine Learning and Teaching, and Human-Centered Machine Learning. He is part of the HUE group at Research at Redmond, where he looks at ways to give people rich agency in the ways they participate in human-AI collaboration systems. You can find out more about him at https://www.linkedin.com/in/gonzaloramos/

Justin D. Weisz is a Research Manager at IBM Research in Yorktown Heights, NY. He leads the Human-AI Collaboration team, whose mission is to design, build, and rigorously investigate new forms of human-AI partnerships that enhance and extend human capabilities. He was a co-organizer of the Human-AI Co-Creation with Generative Models (HAI-GEN) workshops at IUI in 2021 & 2022 and is the PI of a project that explores enterprise use cases of generative AI technologies and explainability needs for generative models.

Hal Daumé III 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. His work develops new learning algorithms for prototypical problems that arise in the context of NLP and AI, with a focus 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 a professor for Human-Machine Interaction and Virtual Reality at Technische Hochschule Ingolstadt (THI) with co-affiliation at the CARISSMA Institute of Automated Driving. He is a program manager for User Experience Design and leads the UX/usability research and driving simulator labs. His research interests include HF/ergonomics, adaptive UIs, driver state assessment, and trust/acceptance/ethics in mobility applications. 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

Explainability is an essential pillar of Responsible AI. Explanations can improve real-world efficacy, provide harm mitigation levers, and can 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. In addition to papers, we will host a video track in 2023 (*"The Sanities & Insanities of XAI: Provocations & Evocations"*). Participants submit 90-second videos with provocative content (for example, design fiction, speculative design, or other creative ideas) discussing the future of XAI and human-AI interactions. 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] AccessSIGCHI. 2021. AccessSIGCHI statement on Chi 2022: Virtual or hybrid. https://docs.google.com/document/d/1mMmCXf1HT_7SNlcNpSS7U466WfVzuvRueh8HSSOOUWk/edit

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[4] 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.

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

[6] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, New York, NY, USA, 1591–1602.

[7] 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.

[8] 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]

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

[10] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence (Lecture Notes in Computer Science)*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[11] 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.

[12] 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.

[13] 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.

[14] 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).

[15] 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.

[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] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[18] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (April 2020), 1–15. https://doi.org/10.1145/3313831.3376590 arXiv:2001.02478

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

[20] 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

[21] 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.

[22] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[23] 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.

[24] Simone Stumpf, Adrian Bussone, and Dympna O'sullivan. 2016. Explanations Considered Harmful? User Interactions with Machine Learning Systems. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.

[25] 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.