Human-Centered Explainable AI (HCXAI): Re-examining XAI in the Era of Agentic AI

Upol Ehsan
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA
u.ehsan@northeastern.edu

Amal Alabdulkarim
Georgia Institute of Technology
Atlanta, Georgia, USA
amal@gatech.edu

Kenneth Holstein
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kjholste@cs.cmu.edu

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

Andreas Riener
Human-Computer Interaction Group
Technische Hochschule Ingolstadt
Ingolstadt, Bavaria, Germany
andreas.riener@thi.de

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

Abstract

Making AI explainable requires *more* than algorithmic transparency: it demands understanding *who* needs explanations and *why*. In our *sixth* CHI workshop on Human-Centered XAI (HCXAI), we shift focus to agentic AI systems. LLM-based agents foundationally challenge existing explainability paradigms. Unlike traditional AI that produces single outputs, agents plan multi-step strategies, invoke tools with real-world consequences, and coordinate with other systems; yet current XAI approaches fail to address these complexities. Users need to understand not just *what* an agent might do, but the cascade of actions it could trigger, the risks involved, and why responses take time. Even our expanded HCXAI frameworks struggle with these new demands. Through our workshop series, we have built a strong community making important conceptual, methodological, and technical impact. This year, we re-examine what human-centered explainable AI means in the agentic era, bringing together researchers and practitioners to shape explainability for both users and developers of these systems.

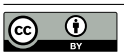
CCS Concepts

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

Keywords

Explainable AI, Responsible AI, Agentic AI, Human-centered AI

ACM Reference Format:
Upol Ehsan, Amal Alabdulkarim, Kenneth Holstein, Min Kyung Lee, Andreas Riener, and Justin D. Weisz. 2026. Human-Centered Explainable AI (HCXAI): Re-examining XAI in the Era of Agentic AI. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3778728



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI EA '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3778728

1 Motivation

Explainable AI (XAI) is not a new field; its roots trace back to the 1980s with expert systems [25]. It experienced a resurgence around 2016 with the DARPA XAI Challenge [18]. XAI, as a domain, is pluralistic [11]: a field concerned with developing techniques, concepts, and processes that can help stakeholders understand the reasons behind the AI system's decision-making [12, 20]. Even compared to the state of the field in 2021, when we ran the first installment of the Human-Centered Explainable AI (HCXAI) workshop at CHI [14], the XAI landscape has substantially matured. For instance, techniques that were experimental at the time have now been productized, such as IBM's AI Explainability 360 Toolkit[1], Microsoft's InterpretML[2], and Google's Vertex Explainable AI[3].

Similarly, the regulatory landscape also shows a demand for XAI to foster Responsible AI (RAI). For instance, the EU AI Act and the US AI Bill of Rights both call for making AI systems explainable. Moreover, the advent of large language models (LLMs) has ushered in a new set of problems not previously encountered in XAI (e.g., hallucinations/confabulations), as well as new needs to explain the behaviors of physical systems like automated vehicles [27] and robots [24].

Today, we have entered an era in which LLMs are now being used to power AI agents: entities capable not only of conversing in natural language with a human user, but also of using tools to collect information or execute actions in external systems. These agents can produce "chain of thought" (CoT) reasoning, but recent work questions whether CoT rationales can be treated as explanations [4].

In increasingly shortening epochs, *the design material [9] of AI continues to evolve*, from statistical (and inherently interpretable) models to neural networks, then to LLMs, and now to LLM-based AI agents. This evolution demands a re-examination of the central assumptions and expectations of the field. Thus, we observe a critical need to continue hosting a community of HCI and AI scholars to identify the grand challenges of explainable agentic AI at CHI 2026.

[1] https://github.com/Trusted-AI/AIX360
[2] https://interpret.ml
[3] https://cloud.google.com/vertex-ai/docs/explainable-ai/overview

Ever since Ehsan and Riedl [11]'s work coined the term and charted the vision, the field of human-centered XAI (HCXAI) has also matured in tandem with the broader discourse in XAI and responsible AI (RAI). HCXAI encapsulates the philosophy that not everything that is important to understanding an AI system lies within the "black box" of an AI model. Crucially, important insights about the system often lie outside of 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]. Given that real-world AI systems exist in sociotechnical settings, algorithmic transparency alone is insufficient at making AI systems explainable [15]. Increasing the aperture of XAI to include sociotechnical and human factors helps 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 HCXAI's focus on understanding and designing for human needs, CHI continues to be the most appropriate venue to address the human side of XAI.

With five consecutive installments, the HCXAI workshop series is one of the longest running and most well-attended one at CHI – over 450 researchers, practitioners, policy makers from over 21 countries spanning 9 sectors have found community with us since 2021. We propose the **sixth HCXAI installment** for *intellectual* and *community* reasons. From an *intellectual* angle, the AI landscape has significantly shifted with the advent of LLM-based agents and peoples' needs for explainability are unclear in this new paradigm. This observation was made at our 2025 workshop by Ferrario [17], who asserted:

> *"...it is the HCXAI framework that can help address[] the rise of AI agents [by] providing appropriate explanations for their activities, ensuring accountability across multi-layered and multi-system interactions, and maintaining trust in settings where the interaction between humans and digital resources... are mediated by AI agents."*

Ferrario [17] argued that the rise of AI agents creates a "triadic relation among the human user, their AI agent(s), and the digital resources accessed by them." We push this view even further by considering both:

- The **users** of LLM-based AI agents, who seek to use them to complete meaningful personal and work tasks, and
- The **developers** of LLM-based AI agents, who seek to understand, shape, and debug the agent's behaviors.

The explainability needs for these two groups remain unclear. In another paper presented at our workshop in 2025, Manuvinakurike et al. [22] found that chain-of-thought reasoning traces "tend to produce explanations without explainability, in that they do not improve the ability of end users to better understand systems or achieve their goals." Thus, we need to develop a new understanding of exactly what users' and developers' explainability needs *are*, in addition to new mechanisms to *produce* those explanations.

A *compelling community need* has also emerged. CHI's remarkable expansion in recent years presents both opportunity and challenge – though the conference now welcomes a broader spectrum of voices, the proliferation of sessions, talks, and publications risks overwhelming newcomers, *particularly early-career researchers*. Creating anchoring points and fostering genuine connection becomes essential if we expect emerging scholars to embrace CHI as their intellectual home and sustain their engagement across years. In the past twelve months alone, we've fielded more than 63 earnest requests from past attendees insisting their "core [HCXAI] community must persist" (Workshop Attendee, XAI Scholar, North America). One participant crystallized this collective sentiment:

> *Been a regular since 2022, and this workshop has become central to my research practice...It's not just the content but the people here. Somehow the organizers designed it in a way that preserves authentic connection even as attendance soars past 100... the continuity of this community matters deeply to me* (HCI Researcher, Southeast Asia)

The 6^{th} workshop iteration draws from a rich five-year trajectory spanning 2021-2025 [13–16]. Our opening year grappled with establishing rigorous evaluation frameworks for XAI systems. The 2022 gathering welcomed broader perspectives, shifting focus toward recognizing and mitigating potential XAI harms. By 2023, an "infrastructural lens" emerged, probing how "seamful information" [10] across AI development phases could enhance explanatory power. In 2024, there was a critical reflection of whether "opening the black box" is still a realistic goal for XAI in the era of LLMs. In 2025, participatory design and sociotechnical assessment of LLM explanations took center stage. This 2026 workshop maintains our position bridging disciplinary boundaries, engaging HCXAI through *theoretical, methodological,* and *implementation* perspectives while *interrogating XAI amid the rise of agentic AI*. We pursue dual aims: (1) deepen critical yet constructive engagement with *operationalizing* human-centered XAI and (2) nurture an inclusive community spanning diverse HCXAI stakeholders.

The following list of guiding questions is *not* an exhaustive one; rather, it is provided as a source of inspiration of our workshop:

- What are users' and developers' explainability needs for LLM-based AI agents? What do they need to know about an agent, both before and during interaction, to use it effectively? How does this information shape their trust and reliance on the agents?
- What kinds of explainability can foster user trust and accountability, and provide an adequate level of control over their data privacy?
- To what extent do chain-of-thought traces provide meaningful information to end users? Do they contain useful insights that help users interact more effectively with the agent, or are they merely information overload?
- How do developers use an agent's reasoning and tool-calling traces to debug an agent's behavior? How do they calibrate their trust in this information?
- How do agent designers & developers make decisions about what information to show end users about the reasoning and tool-calling behaviors of their agent?
- How should explanations of multi-step agent plans balance detail and simplicity, and when do users need complete reasoning versus just the goal?

- What empirical methods can assess whether explanations support *appropriate* trust calibration (trusting agents when warranted and remaining skeptical when not) rather than merely increasing trust?

2 Human-centered XAI: Background and Opportunities

LLM-based agents fundamentally differ from static ML models in explainability needs. Agents maintain memory, create multi-step plans, use external tools with real-world effects, and coordinate with other agents. Yet, recent work reveals [2, 5, 8, 21, 26] that chain-of-thought explanations are systematically unfaithful: models rationalize pre-selected answers, hide reasoning gaps, and omit computational shortcuts from their traces. This creates urgent HCXAI issues that we aim to address in this workshop.

Over the past several years, we have built a richly cross-functional community largely through our use of inclusive, social-science-informed analytic frameworks. Science and Technology Studies provides the lens of "Social Construction of Technology" (SCOT) [7], a framework that recognizes that the function and operation of technology emerges from the interplay between technologies, humans, and organizational/societal-level systems. Using such a lens enables us to allocate attention to actors working from disparate positions in this interplay. For example, one motivating principle across the history of HCXAI workshops has been to examine the "who" of explainability: how diverse *relevant social groups* (including stakeholders such as researchers, developers, or policymakers) draw on explainability's *interpretive flexibility* (meaning, the flexibility to support multiple concurrent diverging interpretations). This work takes on ever-greater urgency, as more stakeholders are impacted by, and participate in, the construction of what explainability means, how it is defined, for whom, and for what. For example, terms such as explainability, interpretability, and transparency, have been used interchangeably across distinct communities [1, 3, 23]. In addition to the questions around "by who?" and "for whom?" we ask also *why* [11] explanations are built, as explanations are deployed to serve a range of purposes including *causality, fairness, accessibility, trustworthiness, and interactivity* [3].

Any effort towards deriving value from explanations demands identifying and addressing ambiguities around definitions, audiences, and evaluation methods. This approach enables us to make drive the field and its outputs forward at three levels – conceptual, methodological, and technical.

3 Organizers

The Organizing Committee is uniquely positioned to execute the visions of our workshop. We are a global spanning industry and academia and bridges relevant XAI disciplines including AI, HCI, Sociology, and Psychology. Beyond hosting previous versions of this workshop, the committee has a combined experience of organizing over 30 workshops and conferences [6]. We will use our experience to avoid common pitfalls and leverage the strengths of both worlds.

Upol Ehsan is a Harvard Berkman Klein Fellow and Assistant Professor of Computer Science at Northeastern. His work pioneered the domain of Human-centered XAI, receiving multiple awards at ACM CHI and HCII. Alongside serving as ACs in HCI and AI conferences (e.g., DIS, IUI, NeurIPS) and being the lead editor of the TiiS Journal Issue on HCXAI, he has been the lead organizer for HCXAI workshop series at CHI from 2021-2025. He has served on national AI task forces in the U.S., EU, and South Asia, and as expert advisor to UK's CDAC SAFE AI consortium informing national strategies on AI governance.

Amal Alabdulkarim is a doctoral candidate in Computer Science at Georgia Tech, working at the intersection of reinforcement learning, narrative generation, and explainable AI. Her research aims to make complex, sequential decision-making systems more transparent and human-understandable especially for users without deep AI expertise.

Ken Holstein is an Assistant Professor at Carnegie Mellon University's Human-Computer Interaction Institute. His research explores how AI systems can effectively augment and bring out the best of human expertise in real-world design and decision-making. This research has been recognized with paper awards at ACM CHI. Ken has co-organized numerous workshops at ACM CHI, CSCW, and FAccT.

Min Kyung Lee is an Assistant Professor and a director of the Human-AI Interaction lab in the School of Information at the University of Texas at Austin. Her current research investigates enabling participatory AI to create empowering and effective workplaces and cities that prioritize the well-being and shared growth of workers and communities. Dr. Lee has organized several workshops and panels at the ACM CHI and CSCW, and is a steering committee member for the ACM FAccT.

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. He leads the HCIG research group. His research interests include HF/ergonomics, adaptive, explainable UIs, driver state assessment, and trust, acceptance, and 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).

Justin D. Weisz is a Senior Research Scientist and Manager at IBM Research in Yorktown Heights, NY. He co-leads the Human-Centered Trustworthy AI strategy and his work explores the design of useful, safe, and personal AI experiences in enterprise settings. He has co-organized numerous workshops including HAI-GEN at IUI (2021-2025), HCXAI (2023) and ToMinHAI (2024) at CHI, and ToMinCUI (2025) at CUI. Dr. Weisz is an ACM Senior Member.

3.1 Program committee

In addition to the organizing committee, we will recruit a program committee to review submitted papers. Based on past data, our workshop has received approximately 50-60 submissions, and we have a successful track record in recruiting dedicated PC members who provide thoughtful and constructive reviews to our authors.

4 Workshop Logistics

4.1 Plans to publish proceedings

Proceedings from our workshop will be published on our web site, joining the proceedings from previous years: http://hcxai.jimdosite.com. We also plan to publish our proceedings on the EU open

research platform Zenodo[4], as in previous years. In terms of **offline materials**, participants who might miss certain sessions can catch up on proceedings published on our website and Zenodo.

4.2 Workshop length and expected size of attendance

Our workshop will be **two sessions** in duration. On average, the expected size of our workshop has been 40-60 participants (when the event was hybrid, the attendance was between 90-140 participants).

4.3 Pre-Workshop Plans

Our pre-workshop plans serve two goals: **advertising** (to raise awareness and receive strong submissions) and **building community**. 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, Sociology, Psychology, and Policy. We will primarily use Twitter and LinkedIn to advertise, leveraging our network of over 40,000 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. Building on successful practices from previous years, we will also host optional online meetups and presentations in the month leading up to the workshop. In the past workshops, these sessions helped newcomers integrate into the HCXAI community, provided student presenters with opportunities to practice their talks, and built momentum for the in-person event.

4.4 Workshop activities

The primary goal of our workshop will be to **define a research agenda** for the next 2-3 years of human-centered explainable AI research. To ensure we include the right set of people who can craft this agenda – including academics, students, and industry researchers – we will advertise our Call for Participation in relevant places including the `CHI-announce` and `IUI-announce` mailing lists, on the human-centered AI Slack workspace, on social media, and within our own personal networks.

During the workshop, we will engage our participants with three activities – a **keynote speaker**, a series of **Spotlight talks and discussions**, and a **design futures activity** – which we show in a tentative schedule in Table 1.

As in previous years, we will set up a **Discord server** that serves as an online space for discussions before, during, and after the workshop. In the past, participants have used Discord to queue their questions during talks, catch up on missed presentations, and directly engage with other workshop participants, both during and after the workshop's conclusion.

Session 1 (90m)	
5m	Introduction of workshop organizers, topics, and goals
45m	Activity 1: Keynote speaker + Q&A
40m	Activity 2: Spotlight talks + Q&A
Session 2 (90m)	
70m	Activity 3: Design futures activity
20m	Wrap up: Reflections & next steps

Table 1: Tentative workshop schedule

4.4.1 Activity 1: Keynote Speakers. We will begin with a panel of keynote speakers – what we have historically called the **main event** – who will help set the context for the workshop by discussing real-world experiences in designing & deploying agentic AI. This event is run as a moderated discussion between one of the workshop organizers and the panel, and it includes a high degree of interactivity with workshop attendees as the moderator collects and asks their questions throughout the discussion.

We will recruit speakers across industry, academia, and public policy who span the theory-to-practice spectrum of XAI and agentic systems. In past workshops, we have recruited high-profile researchers and practitioners for this event, including Mary Gray (Senior Principal Researcher, MSR), Maryam Ashoori (VP of Product Management, IBM), Jacob Metcalf (Program Director, Data & Society), Kush Varshney (IBM Fellow, IBM Research), Tim Miller (a seminal figure in XAI, University of Melbourne), Tania Lombrozo (a seminal figure in the study of explanations, Princeton University), Andres Paez (Professor, University of the Andes) and more. This event will ground our discussions in current, real-world practices for developing agentic AI systems.

4.4.2 Activity 2: Spotlight Talks. We will invite accepted authors to present their work in Spotlight talks that highlight explainability needs for users and developers or techniques for addressing those needs, coupled with discussions to elicit a set of requirements for the explainability of agentic AI.

4.4.3 Activity 3: Design Futures Activity. Experiential learning forms the core of HCXAI workshops. Previously, we used design futures activities [19] where participants operationalized Seamful XAI [10] by anticipating AI system breakdowns and leveraged seams to reveal AI blind spots. These sessions consistently generated high participant engagement.

This year expands beyond traditional activities. Two industry experts who have deployed agentic systems at scale have confirmed participation, grounding the workshop in real-world use cases. Participants will work in groups to create vignettes illustrating how agentic AI explainability requirements apply across domains: software engineering, design, product management, scientific discovery, and HR. These vignettes will guide HCXAI research directions in coming years.

The activity unfolds on Mural, beginning with *problem framing* where industry experts and organizers present the challenge. Participants then move to *current state assessment*, reflecting on existing

[4]https://zenodo.org/

HCXAI approaches for agentic AI and identifying improvement areas. The core work happens during *future envisioning*, where groups develop two contrasting scenarios: a practical future that meets agentic explainability requirements in their chosen domain, and a cautionary scenario where explanations become weaponized. The session concludes with *feasibility analysis*, where groups evaluate interventions to promote positive outcomes while preventing negative ones, identifying challenges and developing mitigation strategies. These resulting vignettes will provide concrete research directions for the HCXAI community.

4.5 Post-Workshop Plans

As in previous years, post-workshop, we will publish accepted papers on our website and the EU open research platform Zenodo and write a summary article for the Human-Centered AI blog on Medium[5], especially focused on the outcomes of the design futures exercise. This year, we will adopt a participatory approach where participants are requested to co-write sections of the wrap-up report. This democratizes how we synthesize workshop insights. Community building will continue through our Discord server, which has proven a fertile ground for sustained collaboration—past participants have co-authored journal submission. One notable post-workshop outcome in the past was the synthesis of the first dedicated HCXAI special issue in ACM Transactions on Interactive Intelligent Systems (TiiS). These post-workshop efforts ensure insights from the event shape the field well beyond the conference.

5 Accessibility

We are committed to fostering an inclusive workshop environment that welcomes diverse perspectives and enables equitable participation. Our organizing committee intentionally bridges academia and industry to bring varied institutional perspectives and intersectional identities that bring lived experiences to accessibility needs. To support equitable dialogue, we use Discord as our primary discussion platform, allowing all participants to contribute without conversations being dominated by any single voice. For the workshop itself, we prioritize accessibility in three key areas. First, submitted position papers and posters will follow SIGCHI Accessibility Guidelines to ensure materials are perceivable and understandable by all participants. Second, our collaborative activities will use Mural, which offers built-in accessibility features including keyboard navigation, screen reader compatibility, and adjustable zoom. We will publicize our tool choices in advance and run accessibility checks on all workshop boards before the event. Third, we encourage participants to contact organizers with specific accommodation needs or suggestions for improving accessibility. Our goal is to create an environment where all attendees can fully engage with workshop content and contribute to shaping the research agenda for human-centered explainable AI.

6 Call for Participation

Explainable AI (XAI) 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. This workshop serves as a junction for cross-disciplinary XAI stakeholders to examine how human-centered perspectives in XAI can be operationalized at the conceptual, methodological, and technical levels. We invite *position or empirical papers* making justifiable arguments 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).

In the *sixth* Human-Centered Explainable AI (HCXAI) workshop, we are focused on reinterperting explainability in the context of LLM-based AI agents. We specifically seek work that examines explainability needs and methods for both *users* and *developers* of agentic AI systems through critical examinations, formative or evaluative empirical studies, new prototypes, and/or design fictions.

Papers should follow the CHI Extended Abstract format (up to 4 pages, excluding references) and be submitted through the workshop's submission site (https://hcxai.jimdosite.com/). Papers should not be anonymized as reviews will be single-blind. Accepted papers will be presented at the workshop, which will include a keynote and a design futures activity to shape a research agenda for the future of HCXAI research.

[5]https://medium.com/human-centered-ai

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] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. arXiv:2503.08679 [cs.AI]

[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] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v1.

[5] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-Thought Is Not Explainability. (2025). https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Cot_Is_Not_Explainability.pdf Preprint. Under review..

[6] Zoe M. Becerra, Nadia Fereydooni, Andrew L. Kun, Angus McKerral, Andreas Riener, Clemens Schartmüller, Bruce N. Walker, and Philipp Wintersberger. 2021. Interactive Workshops in a Pandemic: The Real Benefits of Virtual Spaces. *IEEE Pervasive Computing* 20, 1 (2021), 35–39. doi:10.1109/MPRV.2020.3044072

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

[8] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410* (2025). https://arxiv.org/abs/2505.05410 Alignment Science Team, Anthropic.

[9] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), 278–288. doi:10.1145/3025453.3025739

[10] 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).

[11] Upol Ehsan and Mark O Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach, In International Conference on Human-Computer Interaction. *arXiv preprint arXiv:2002.01092*, 449–466.

[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, Elizabeth A Watkins, Philipp Wintersberger, Carina Manger, Sunnie SY 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*. 1–6.

[14] 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.

[15] 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.

[16] 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.

[17] Andrea Ferrario. 2025. Expanding HCXAI in the Age of AI Agents: Challenges and Recommendations. *Proceedings of the 5th Workshop on Human-Centered Explainable AI, held in conjunction with CHI 2025, Yokohama, Japan* (2025). doi:10.5281/zenodo.15170379

[18] D Gunning. 2016. Explainable artificial intelligence (XAI). Defense Advanced Research Projects agency.

[19] R. Jungk and N.R. Müllert. 1987. *Future Workshops: How to Create Desirable Futures*. Institute for Social Inventions. https://books.google.com/books?id=-usQAgAACAAJ

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

[21] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread* (2025). https://transformer-circuits.pub/2025/attribution-graphs/biology.html

[22] Ramesh Manuvinakurike, Emanuel Moss, Elizabeth Anne Watkins, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2025. Thoughts without Thinking: Reconsidering the Explanatory Value of Chain-of-Thought Reasoning in LLMs through Agentic Pipelines. *arXiv preprint arXiv:2505.00875* (2025).

[23] 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. arXiv:1811.11839 doi:10.1145/3387166

[24] Paul Schmitt, Swapna Joshi, Bruce N Walker, and Azra Habibovic. 2024. Symbiotic-A Workshop on Transferability between Human-Robot Interaction (HRI) and Autonomous Vehicle Interaction Research. In *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 273–276.

[25] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.

[26] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 74952–74965. arXiv:2305.04388 [cs.CL] https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

[27] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 252–261. doi:10.1145/3409120.3410659