AI CHAOS! 2nd Workshop on the Challenges for Human Oversight of AI Systems

Malik Khadar
Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota, USA
khada005@umn.edu

Julia Cecil
Department of Psychology
LMU Munich
Munich, Germany
julia.cecil@psy.lmu.de

Leon Van Der Neut
Delft University of Technology
Delft, Netherlands
l.m.b.vanderneut@tudelft.nl

Nikola Banovic
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan, USA
nbanovic@umich.edu

Kevin Baum
Center for European Research in Trusted AI (CERTAIN)
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Saarland, Germany
kevin.baum@dfki.de

Stevie Chancellor
Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
steviec@umn.edu

Enrico Costanza
UCL Interaction Centre
University College London
London, United Kingdom
e.costanza@ucl.ac.uk

Motahhare Eslami
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
meslami@andrew.cmu.edu

Anna Maria Feit
Saarland Informatics Campus
Saarland University
Saarbrücken, Germany
feit@cs.uni-saarland.de

Susanne Gaube
Global Business School for Health (GBSH)
University College London (UCL)
London, United Kingdom
susanne.gaube@ucl.ac.uk

Ujwal Gadiraju
Web Information Systems
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

Harmanpreet Kaur
University of Minnesota
Minneapolis, Minnesota, USA
harmank@umn.edu

Abstract

As AI systems are increasingly adopted in high-stakes domains such as healthcare, autonomous driving, and criminal justice, their failures may threaten human safety and rights. *Human oversight* of AI systems is therefore critically important as a potential safeguard to prevent harmful consequences in high-risk AI applications. The global regulatory and policy landscape for AI governance remains understandably fragmented and diverse. While frameworks like the European AI Act require human oversight for high-risk AI systems, there is currently a lack of well-defined methodologies and conceptual clarity to operationalize such oversight effectively. Independent of policy and regulation, poorly designed oversight can create dangerous illusions of safety while obscuring accountability. This interdisciplinary workshop aims to bring together researchers from various disciplines, including AI, HCI, psychology, law, and policy, to address this critical gap. We will explore the following questions: (1) What are the greatest challenges to achieving effective human oversight of AI systems? (2) How can we design AI systems that enable meaningful human oversight? (3) How do we assign responsibilities to and support the various stakeholders involved in oversight? Through talks and interactive group discussions, participants will identify oversight challenges; examine stakeholder roles; discuss supporting tools, methods, and regulatory frameworks; and establish a collaborative research agenda. Our central goal is to further a roadmap that enables effective human oversight for the responsible deployment of AI in society.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3778736

Keywords

Human Oversight of AI, Explainable AI, Human-AI Interaction, Human-centered AI

ACM Reference Format:
Malik Khadar, Julia Cecil, Leon Van Der Neut, Nikola Banovic, Kevin Baum, Stevie Chancellor, Enrico Costanza, Motahhare Eslami, Anna Maria Feit, Susanne Gaube, Ujwal Gadiraju, and Harmanpreet Kaur. 2026. AI CHAOS! 2nd Workshop on the Challenges for Human Oversight of AI Systems. 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/10.1145/3772363.3778736

1 Motivation

In recent years, we have witnessed the increasing deployment of AI-enabled applications in high-stakes domains, such as medical diagnosis and treatment recommendations, autonomous driving systems, predictive policing, recidivism assessment, and automated hiring processes [2, 13]. Failures of applying AI in these contexts—be it at the level of conceptualizing the application or post-deployment—directly threaten human safety and fundamental human rights [14, 17]. The opacity and complexity of many AI systems further compound these risks [4, 18]. However, stakeholders often cannot effectively oversee automated decisions that significantly impact human lives (e.g., through monitoring or detecting failures or harmful outcomes). This reinforces the growing call for human oversight in emerging ethical guidelines and legislation (such as the European AI Act [11]), especially for AI use in high-risk contexts.

We consider "**Human Oversight**" as an umbrella term that encompasses efforts towards similarly important desiderata of human-AI collaboration, including responsible AI, trustworthy AI, meaningful AI, explainable AI, fair and ethical AI, etc. [4, 7, 8, 12]. Following Sterz et al. [20], we understand human oversight as the supervision of an AI system by at least one natural person, who is systematically prepared and empowered to consciously monitor its operations and intervene, if necessary, in order to substantially reduce AI-induced risks. Although human oversight of AI systems is a potentially effective strategy to mitigate the risks of AI bias and errors, it demands that humans are capable of monitoring AI systems effectively, recognizing critical situations or AI behavior, and executing timely interventions at scale [1, 19]—a task far from trivial [6]. Ultimately, this makes oversight a core challenge for HCI: 1) to decide if human oversight is not only feasible, but also desirable; and 2) if yes, how to design socio-technical systems that render oversight reliable and meaningful in practice.

Currently, the discourse on human oversight lacks conceptual clarity and methodologies to understand and implement effective human oversight in AI-powered systems [20]. Poorly designed oversight mechanisms can, for example, create a false sense of safety, leading stakeholders to believe that systems are properly monitored when they are not [3, 6, 10]. They can obscure accountability and delay the identification of system problems [15, 16]. They can also affect cognitive, affective, and motivational processes [5, 9], in particular when domain experts change their role from making decisions themselves to overseeing automated systems. Therefore, effective human oversight requires not only technical innovation but also a deep understanding of human behavior and the affordances of socio-technical systems.

We propose this workshop as an important step within an emerging conversation on the topic of human oversight, with the goal of adding conceptual and design clarity, and methodological rigor. Originating from a recent Dagstuhl seminar,[1] researchers started engaging in interdisciplinary discussions to address these fundamental challenges, including an upcoming workshop at the ACM conference on Intelligent User Interfaces (IUI) in March 2026. While these efforts focused on conceptual clarity and interface design respectively, our proposal for CHI aims to bridge communities and create interdisciplinary collaboration among researchers across fields such as AI, formal methods, HCI, psychology, law, and policy. The CHI community presents a unique platform to foster these important and timely objectives in one place, as it brings together researchers and practitioners from diverse disciplinary backgrounds and methodological traditions, with a longstanding track record of advancing human-centered approaches to the design, deployment, and evaluation of interactive systems. This makes CHI an ideal venue to synthesize emerging insights, stimulate new collaborations, and ensure that conceptual developments around human oversight translate into actionable design principles and practical tools. As such, we anticipate outcomes from this CHI workshop to be widely-applicable and interdisciplinary by design. Subsets of organizers will then bring this conversation to various relevant communities (e.g., similar workshop proposals are planned for additional conferences, including FAccT, HCOMP, UMAP, and AAAI); and we aim to compile definitions, examples, and solutions for human oversight of AI systems through these collective, interdisciplinary efforts initiated at a bigger scale at CHI.

In two 90-minute sessions, our workshop will first solicit perspectives on current challenges in human oversight of AI systems, and then collectively work on designing solutions to achieve effective human oversight. Our call for participation will specifically ask for participants' unique understanding of human oversight and examples of good/bad oversight. We will analyze these participant submissions to schedule lightning talks in the first 90-minute session: a context-building activity where people can gain clarity on what our community understands as "oversight." Using this shared context, the second 90-minute session will focus on group discussions to design effective oversight in human-AI collaborative settings. Organizers will use these activities to generate bottom-up definitions and examples of oversight, as well as design desiderata for effective oversight systems.

Our focus and outcomes will include, but not be limited to, the following topics:

[1] https://www.dagstuhl.de/en/seminars/seminar-calendar/seminar-details/25272, https://www.dagstuhl.de/en/institute/news/2025/laesst-sich-ki-beaufsichtigen

Core HCI Questions for Human Oversight of AI Systems

To enable effective human oversight of autonomous AI systems requires interdisciplinary work on research questions that are fundamentally about human-computer interaction, such as:
(1) *How should automated AI systems communicate their reasoning, uncertainty, or limitations in ways that humans can understand, evaluate, and act upon?*
(2) *How can interfaces be designed to give overseers meaningful, timely, safe, and context-appropriate opportunities to intervene?*
(3) *How can oversight mechanisms help humans form accurate mental models of system reliability without fostering over- or under-reliance?*
(4) *How can oversight mechanisms include humans in a meaningful way that supports their motivation, autonomy, and job satisfaction?*
(5) *How are oversight responsibilities and accountability assigned and how are they made transparent?*
(6) *Finally, more broadly, how do constructs of power, governance, and legality intersect with various stakeholders and mechanisms of oversight?*

Figure 1: A selection of HCI questions that are critical to human oversight of AI systems.

- Identifying current challenges in effective human oversight design for AI systems.
- Understanding the roles and responsibilities of various stakeholders in human oversight.
- Sharing perspectives about human oversight, identifying and addressing open challenges.
- Designing methods, tools, processes, and resources to support stakeholders in overseeing AI systems.
- Identifying the user interface-related opportunities and challenges of human oversight.
- Exploring the trade-off between effective human oversight for risk mitigation and challenges of human autonomy, development, and self-efficacy.

2 Call for Participation

AI systems are rapidly becoming part of everyday life, from healthcare to autonomous driving to job hiring. But when things go wrong, who is really in charge? Human oversight is supposed to keep AI in check, yet we do not have clear answers on what effective oversight means, how to design for it, scale it up, or make it meaningful. Poorly designed oversight can create illusions of safety while obscuring responsibility, whereas well-designed approaches can empower humans to identify critical moments and act effectively.

In this workshop on human oversight of AI systems, we invite contributions from anyone interested in the design, evaluation, or governance of oversight practices in AI systems. We request submissions via the linked Google Form, which solicits participants' understanding of oversight, perspectives on what should and should not be an "oversight" task, and examples of poor and effective oversight that they can speak to. Answers to these questions will guide our program to be suited to our shared understanding of human oversight. We particularly encourage submissions from participants from diverse disciplines to enable rich interdisciplinary conversations on this topic.

3 Pre-Workshop Plans

We will advertise the workshop in research mailing lists and via social media. The promotional materials will include links to our workshop website. Prospective participants will be invited to complete a short survey designed to capture their experiences and opinions on specific instances of oversight. This will provide valuable data in the form of expert perspectives on a variety of oversight contexts, and will directly inform and shape the design of our planned workshop activities.

The organizing committee will review and select submissions in a manner that captures a wide range of perspectives on oversight of AI, prioritizing interdisciplinarity and submissions that describe unique contexts of oversight. Two weeks before the workshop begins, the organizers will meet to select a subset of submissions; we will invite their authors to provide lightning talks on the examples of oversight they submitted (ideally ~10 examples), selecting from the remaining set of submissions and authors if anyone declines. We expect 50 participants to enable groups of 5 for discussing each of the 10 oversight examples solicited before the workshop for lightning talks.

4 Workshop Structure and Activities

The workshop is planned as two 90-minute sessions with a break in between. Table 1 outlines the key activities.

Session 1 will begin with brief opening remarks to align participants with the workshop goals (with a focus on defining oversight of AI systems as a boundary object to connect diverse disciplines), and introduce the workshop activities and the organizers. This will be followed by a series of lightning talks that occupy the remainder of this session. These lightning talks will be pre-selected based on workshop participants' submissions to our call. Our goal is to have around 10 diverse oversight examples discussed in these lightning talks. Participants will have access to a shared virtual whiteboard where

they may take notes and express their opinions on the various contexts of oversight described in the lightning talks. We will evaluate several virtual whiteboard applications for accessibility prior to adoption.

During the break between the two sessions, the organizers will review notes taken on the virtual whiteboard to identify themes related to the difficulties of oversight and frame them as design challenges for Session 2.

Session 2 will begin with instructions and group formation for the following activities. There will be a group associated with each of the lightning talks to provide a variety of oversight contexts. At least one organizer will be present in each group. Once all groups are formed, the *design activity* will start, where each group will select a design challenge to tackle as it intersects with their group's oversight context. The goal of the activity is to produce an artifact such as a storyboard, low-fidelity prototype, or Wizard of Oz script, that represents the group's design for effective oversight in the context assigned to the group. Next, there will be a *red-teaming jigsaw activity* involving design critique, where groups view the artifacts of other groups and try to find failure modes, biases, and unintended consequences of their designs. The organizers will not shift groups and instead will stay to help explain the designs. Afterward, groups will return to their designs to debrief and brainstorm how to address the issues brought up during the red-teaming activity. We will conclude the session with closing remarks that remind participants where the outcomes of the workshop can be found and encourage them to pursue ideas and collaborations that began during the workshop.

After the workshop, we will asynchronously *distribute the emails* of consenting attendees if they wish to stay in touch. In the weeks following the workshop, we will synthesize the outcomes of the activities into general guidelines for human oversight of AI systems, including open challenges and research gaps identified by participants. We will describe these guidelines and challenges in a *blog post on the workshop website*, which will also include images of the artifacts that participants produce. Once we publish the blog post, we will share it with all of our participants. To increase visibility, we will organize the contents into an *ACM Interactions submission on oversight problems*, which we will also share with participants and the larger research communities.

5 Accessibility

We will use the survey from our call for participation to anticipate accessibility needs by including questions about accessibility. Based on the responses, we will work closely with the conference organizers to ensure that all required support, resources, and accommodations are in place to meet those accessibility needs. Additionally, when designing our workshop website, we will make sure that it complies with W3C Accessibility Standards. [2]

[2]https://www.w3.org/WAI/standards-guidelines/

6 Organizers

The following workshop organizers span different continents and cultural backgrounds, represent complementary disciplinary expertise, are at varying stages of their careers, and correspond to a diverse set of perspectives on human oversight of AI systems.

Malik Khadar is a PhD student at the University of Minnesota in the GroupLens Research Lab. His work centers on the exploration of design paradigms to promote appropriate reliance on AI tools. He is particularly interested in how motivation and data literacy factor into the use of such tools. He sees the workshop as an opportunity to connect with and contribute to the CHI community.

Julia Cecil is PhD student at the LMU Munich with a background in psychology. Her research investigates psychological factors in human–AI interaction, with a particular focus on high-stakes domains such as healthcare and personnel selection. She is especially interested in how cognitive, motivational, and affective processes shape human-AI interaction. She sees the workshop as an opportunity to foster interdisciplinary exchange.

Leon van der Neut is a PhD student in the Web Information Systems group at Delft University of Technology. He works on the AI value-alignment problem from the perspective of professional practice in the Dutch Government. He graduated from public administration and philosophy of science and technology on the topic of responsible practice when using ML models in executive government. He sees this workshop as a valuable means to connecting with and learning from the CHI community.

Nikola Banovic is an Associate Professor of Computer Science and Engineering at the University of Michigan and an Associate Director of the Michigan Institute for Data & AI in Society (MIDAS). His research focuses on the design and evaluation of explanation mechanisms that help end-users critically reflect on what AI-based systems can and cannot do, and when such systems can be useful; this in turn could help end-users scrutinize such systems in a way that allows them to contest adverse AI decisions and appeal negative outcomes of such decisions. Nikola's work on Explainable AI has been recognized with an NSF CAREER award. Nikola co-organized workshops at the intersection of HCI and AI at CHI 2019 and IUI 2020, and served as a CHI 2025 Workshops co-Chair.

Kevin Baum is a philosopher and computer scientist, currently Senior Researcher and research group leader at the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken, where he co-heads the Center for European Research in Trusted AI (CERTAIN) and leads the research group Responsible AI and Machine Ethics (RAIME). His research focuses on the ethics and governance of AI, with particular emphasis on the preconditions, operationalizations, benefits, and risks of human oversight, reason-based approaches to machine ethics and AI alignment, as well as trustworthiness assessments and responsible decision-making under normative

Planned Workshop Schedule	
Time	**Activity (details in text)**
	Before the workshop
	Invite participants to give lightning talks
	First session of workshop
0:00–0:15	Opening remarks
0:15–1:30	Lightning talks
	Second session of workshop
0:00–0:15	Instructions and Grouping
0:15–0:40	Design activity
0:40–1:10	Jigsaw activity
1:10–1:25	Groups debrief and brainstorm
1:25–1:30	Closing remarks
	After the workshop
	Distribute emails, create community mailing list
	Publish and share blogpost through mailing lists and social media
	Publish and share ACM Interactions article

Table 1: Workshop schedule including two 90-minute sessions with a break in between. Note that the 0:00 in the Time column indicates the start of the associated workshop session.

uncertainty. He has co-organized a recurring interdisciplinary track on responsible and trusted AI, fostering dialogue across philosophy, computer science, law, and related fields. He sees this workshop as an opportunity to connect with and learn from the CHI community.

Stevie Chancellor is an Assistant Professor in Computer Science & Engineering at the University of Minnesota - Twin Cities. She studies building human-centered artificial intelligence for mental health in two areas: AI-driven social media and generative AI. Her interest in oversight relates to intersectional areas with mental health and AI – cognition and decision-making; governance and content moderation; and AI safety in public-use settings. She has organized numerous workshops at CHI, CSCW, and ICWSM, and served as the 2023 CSCW Workshops Co-Chair.

Enrico Costanza is Professor of Human-Computer Interaction at University College London (UCL), and deputy director of the UCL Interaction Centre (UCLIC). His research is focussed on helping people make sense of data and on interaction with AI and autonomous systems. He is driven by the societal importance of a problem-focused research agenda: most of his work addresses applications related to energy and environmental sustainability, and more recently to digital health. Enrico has co-organized workshops on human interaction with autonomous systems at CHI and Ubicomp.

Motahhare Eslami is an Assistant Professor at the School of Computer Science, Human-Computer Interaction Institute, and Software and Societal Systems Department, at Carnegie Mellon University. Motahhare's research goal is to investigate the existing accountability challenges in AI systems and utilize methods such as AI auditing to empower the users of algorithmic systems to make transparent, fair, and informed decisions in interaction with AI systems. Motahhare has been named one of the 100 Brilliant Women in AI Ethics. She has co-organized several workshops on the topic of human-centered AI and human oversight of AI systems including two rounds of workshops at CHI, a CSCW workshop and a SIG, a FAccT CRAFT panel, and an HCOMP workshop.

Anna Maria Feit is an Assistant Professor in Computer Science at Saarland University. She leads the computational interaction group where she and her team work on computational methods for UI optimization and adaptation, which is fundamentally based on the empirical understanding of user's interactive behavior in different contexts and the operationalization of this empirical knowledge. As such, she is particularly interested to explore how intelligent user interfaces can optimally support human oversight of AI systems. Anna is also a PI in the center for perspicuous computing (CPEC) which studies human oversight of cyberphysical- and software systems from an interdisciplinary perspective, spanning HCI, formal methods, AI, psychology, ethics, and law.

Susanne Gaube is an Assistant Professor in Human Factors in Healthcare at University College London (UCL), United Kingdom. Her research focuses on understanding and improving

the interaction between humans and AI systems in healthcare. She investigates how AI-enabled clinical decision support systems shape decision-making processes and how to foster effective human–AI collaboration. More broadly, her work explores how technology can be designed and implemented to ensure safe use and meet the needs of healthcare professionals and patients. Susanne has co-organized several workshops, symposiums, and panel discussions on user-centered implementation requirements for AI systems in healthcare (e.g., at RSNA, GIANT Health, and DGPs).

Ujwal Gadiraju is an Associate Professor in the Faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology, Netherlands. He focuses on advancing human-centered AI through innovative computational techniques and systems that enhance human experiences, align AI systems with human values, promote inclusivity, and foster appropriate human reliance on AI systems. Ujwal has co-led several workshops at the intersection of HCI and AI, including at conferences such as CHI, CSCW, IUI, HCOMP, and UMAP.

Harmanpreet Kaur is an Assistant Professor in the Department of Computer Science and Engineering at the University of Minnesota. Her research areas are human-centered AI, explainability and interpretability, and hybrid intelligence systems. She studies these areas in a variety of domains (e.g., exploratory data analysis, workplace wellbeing, knowledge search and sensemaking), applying methods towards both critically evaluating existing systems on meeting their intended user goals, and designing new human-centered systems. She has organized and participated in several events (e.g., workshops, panels, consortia) on these topics of human-centered AI at conferences such as CHI, CSCW, IUI, HCOMP, FAccT, and KDD.

References

[1] Sebastian Biewer, Kevin Baum, Sarah Sterz, Holger Hermanns, Sven Hetmank, Markus Langer, Anne Lauber-Rönsberg, and Franz Lehr. 2024. Software Doping Analysis for Human Oversight. *Formal Methods in System Design* (2024). doi:10.1007/s10703-024-00445-2

[2] Shreyan Biswas, Ji-Youn Jung, Abhishek Unnam, Kuldeep Yadav, Shreyansh Gupta, and Ujwal Gadiraju. 2024. "Hi. I'm Molly, Your Virtual Interviewer!" Exploring the impact of race and gender in AI-powered virtual interview experiences. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 12. 12–22.

[3] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 307–317.

[4] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[5] Cedric Faas, Richard Bergs, Sarah Sterz, Markus Langer, and Anna Maria Feit. 2024. Give Me a Choice: The Consequences of Restricting Choices Through AI-Support for Perceived Autonomy, Motivational Variables, and Decision Performance. arXiv:2410.07728 [cs.HC] https://arxiv.org/abs/2410.07728

[6] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681.

[7] Kyriakos Kyriakou and Jahna Otterbacher. 2023. In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. *Discover Artificial Intelligence* 3, 1 (2023), 44.

[8] Kyriakos Kyriakou and Jahna Otterbacher. 2024. Modular oversight methodology: a framework to aid ethical alignment of algorithmic creations. *Design Science* 10 (2024), e32.

[9] Markus Langer, Kevin Baum, and Nadine Schlicker. 2024. Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. *Minds and Machines* 35, 1 (2024), 1–30. doi:10.1007/s11023-024-09701-0

[10] Johann Laux. 2024. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & society* 39, 6 (2024), 2853–2866.

[11] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* 18, 1 (2024), 3–32.

[12] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[13] Siddharth Mehrotra, Ujwal Gadiraju, Eva Bittner, Folkert van Delden, Catholijn M. Jonker, and Myrthe L. Tielman. 2025. "Even explanations will not help in trusting [this] fundamentally biased system": A Predictive Policing Case-Study. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. 51–62.

[14] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. arXiv:https://www.science.org/doi/pdf/10.1126/science.aax2342 doi:10.1126/science.aax2342

[15] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[16] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.

[17] Rowena Rodrigues. 2020. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology* 4 (2020), 100005. doi:10.1016/j.jrt.2020.100005

[18] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[19] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[20] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2495–2507.