

Centering Civil Society Voices for Social Justice: Critical Reflections and Ways Forward for AI Practitioners

Steven Vethman¹, Marie Mirsch², Louise Aupetit³, Emma Kallina⁴,

¹Sciences Po Law School, Paris

²RWTH Aachen University

³Women in AI

⁴Women At The Table, Research Centre Trust (UA Ruhr), University of Cambridge

mail@stevenvethman.com, marie.mirsch@gdi.rwth-aachen.de, louise@womeninai.com, emma@womenatthetable.net

Abstract

Civil society organizations (CSOs) have played a central role in defending human rights and advancing social justice—yet, their perspectives remain underexplored in academic literature linking social justice to AI practice. Our study centers CSO voices in a reflection on recent recommendations for AI practitioners striving to work towards social justice. We conducted a CSO community workshop and interviews with CSO advocates. Participants critically reflected on the existing recommendations and collaboratively articulated constructive pathways for action. Our findings highlight the value of CSOs as crucial and pragmatic allies towards social justice: our participants acknowledged structural and organizational barriers, without considering them a reason for inaction. Instead, they invited practitioners to take proactive steps within their sphere of influence. Indeed, CSO priorities highlight the key role of motivated practitioners, emphasizing practices that can be exercised by everyone and everyday: humility, respect for diverse voices and perspectives, and a willingness to see and create openings for strategic, pragmatic actions.

1 Introduction

Technological interventions under the heading of artificial intelligence (AI)¹ have become embedded across services and industries, often with the goal of increased automation, cost savings, and efficiencies (Cobbe et al. 2020). However, such AI adoption has also led to various harmful outcomes and infringements of human rights (for an overview see McGregor 2021). Importantly, it has been shown that such harms unduly affect already marginalized communities, exacerbating social injustices in our societies. Some illustrative examples include cases in which black patients received fewer medical treatments (Obermeyer et al. 2019), black inmates fewer sentences including bail (Angwin et al. 2016), and benefit recipients with an immigration status, lower income, or disabilities were overproportionately labeled as fraudulent (Amnesty International 2024b, 2021; Booth 2024; Eubanks 2018).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Acknowledging the ambiguity of the term “artificial intelligence” in limiting the critical reflection we aim for (Bender and Hanna 2025; Monett and Lewis 2017), we have still chosen to use the term as it aligns with the terminology used by our target audience and reaching them is our primary goal.

Whilst public bodies are in the process of establishing regulatory frameworks, legislative acts, impact assessments, and standards for these technological interventions (European Union 2024; Council of Europe 2024; Gerards et al. 2022), it was *civil society organizations* (CSOs) in the first place who identified and emphasized these harms. Indeed, numerous advocacy campaigns, litigation, and public outreach are led by non-profit and/or CSOs (Amnesty International 2021; AlgorithmWatch and Bertelsmann Stiftung 2020; Meaker 2024; Angwin et al. 2016; Birhane et al. 2024). CSOs, by tradition, advocate for communities most at-risk, stepping in where regulation is not yet established or effective. As such, they play a leading role in defending the rights of marginalized communities and amplify their voices to foster social justice. Despite this, the perspectives from CSOs are currently underexplored in academic literature aimed at the very same goal, i.e. to prevent AI systems that exacerbate or perpetuate social injustices (§2.2).

In order to close this gap—i.e. to learn from the established practices from CSOs in advocating for the intersection of AI and social justice—**this study explores CSO perspectives on recent recommendations for AI practitioners**. The recommendations were derived from recent works that identified key actions for motivated AI practitioners, helping them to explore socio-technical questions and mitigate harmful outcomes beyond technical fixes (i.e. Vethman et al. 2025; Mirsch et al. 2026) (§3.1). To move from a technical to a socio-technical lens, these recommendations are grounded in intersectionality, i.e. the lens or collection of movements for inquiry and action on structural social injustices with eye for power relations and unique lived experiences (all concepts explained in-depth in §2.1). Our stance is in line with these previous works that consider inviting practitioners towards applying a socio-technical lens as a promising path towards social justice, i.e. of a complementary value to aforementioned CSO efforts such as litigation and advocacy campaigns to combat structural injustices (see also Ross 2025; Petterson et al. 2026). *AI practitioners* as used in this work refer to people involved in the creation of AI systems with a certain degree of decision power over how the system manifests concretely, especially focusing on technical practitioners but also expanding to e.g. product managers, strategy consultants, research managers or innovation department leads. More specifically, this work fo-

cuses on AI practitioners that are already working towards more responsible AI systems or are planning to do so: if we invite such *motivated* practitioners to expand their perceived circle of influence and support first steps, allow for mistakes, and make room for disagreement (à la Ross 2025; Edmondson 2018), they can become fundamental partners in taking action against structural injustices and intersectional harms (details in Vethman et al. 2025; Mirsch et al. 2026). To ensure that the recommended practices are applicable and reflect AI practitioners’ experiences, the majority of the recommendations in focus have been co-designed with motivated AI practitioners themselves. Our work expands these recommendations through the perspectives and experiences of the community most proficient in the work of social justice: CSOs. This ensures that recommendations already aimed at motivated practitioners include CSO perspectives, increasing the likelihood that any changes to AI practice have social impacts. This results in the following **research questions**:

RQ1: *Given CSOs longstanding experience with advocating for social justice, what are their critical reflections on recent recommendations for AI practitioners designed to promote social justice?*

RQ2: *Based on these reflections from CSOs, which additional ways forward emerge for AI practitioners?*

We address these questions through a collaborative civil society community workshop, jointly hosted by CSOs and critical scholars from academia. In this interactive workshop (n=19), the participants were invited to critically reflect on the recommendations by Vethman et al. (2025) and Mirsch et al. (2026) (RQ1) as well as to propose additional ways forward (RQ2). To dive deeper into proposed actions (RQ2), we conducted semi-structured interviews with CSO advocates or collaborators (n=10), all experienced in (1) directly collaborating with or teaching (motivated) AI practitioners and (2) experienced on social justice, human rights, or inclusion. These selection criteria ensured that interviewees had the necessary understandings and shared language to support moving critical CSO reflection towards concrete ways forward that are understandable to AI practitioners.

This study contributes to the ongoing effort to facilitate socio-technical, social justice-oriented interventions in AI practice. More specifically, we advance this effort through the following **key contributions**. (1) By demonstrating the valuable input that CSOs provide on recommendations that have been built by academics and practitioners, *we highlight CSO perspectives as a crucial area* for social justice-oriented works. Importantly, CSO input was considerably constructive in promising paths and strategies to inflict positive change. We hope that this motivates researchers and regulators to engage more deeply with CSO voices going forward. (2) By refining existing recommendations towards social justice through the perspectives and lived experiences of CSOs’, we generate a *robust set of both critical reflections and strategies for action* that can orient motivated AI practitioners’ practice. These can guide not only industry practice towards a more equal AI future but also be the basis of regulatory efforts or organizational-internal interventions. We outline first directions towards this in §5.4.

2 Key Concepts and Related Work

To ensure clarity, we take time here to create a shared understanding of foundational concepts of this work (that are often brushed over) (§2.1). We then position the contribution within related works on using the lens of intersectionality to expand the algorithmic to the socio-technical frame as well as the role of CSOs therein (§2.2).

2.1 Key Concepts

Social Justice & Intersectionality. There is broad agreement that justice is desirable, even though its concrete interpretation varies considerably across contexts (Campbell and Mancilla 2012). Often the notion refers to a certain (a)symmetry being (un)desirable, e.g. people having similar freedoms or capabilities to exercise their rights (Sen 2008) or equal liberties for all unless this disadvantages those systematically worse off (Rawls 1971). *Social justice*, in general, concerns questions of justice within social structures. We focus in this paper on the specific interpretation of social justice in the term of *structural injustice* according to Young (2011), understood as injustice arising from the normal functioning of social systems rather than from intentional wrongdoing. Mechanisms through which structural injustice manifests include persistent and group-differentiated forms of disadvantage that are produced and sustained by legal systems, markets, political institutions, cultural norms, or technologies. A key implication is that all who participate in these systems share a political responsibility for social justice. Not because they directly cause harm, but because they participate in and benefit from social structures that produce unjust outcomes. Hence, to move towards achieving social justice requires transforming those structures, not merely compensating victims or correcting individual failures (Young 2011). Historical fights against structural injustices in the context of Black feminism (Combahee River Collective 1978; Berger and Guidroz 2009) have led to the rise of *intersectionality* as a lens for centering the harms, historical context, and lived experiences of multiply marginalized people in addressing questions of social justice (Crenshaw 1989; Collins and Bilge 2020). Intersectionality stresses the structural nature of social injustices and the need for both critical analysis (inquiry) and action (praxis). Intersectional approaches particularly elucidate how multiple social categories, such as race, gender, or class, interact to shape lived experiences and how these experiences position people differently within the world, especially within structural systems of power (Xenidis 2022; Crenshaw 1991).

The Algorithmic Frame & The Socio-Technical Frame. AI systems are embedded in these same social structures (Kasirzadeh 2022; Browne et al. 2023; Naudts 2024). As tools of analysis, they bear potential in revealing injustices as a reflective mirror of our society (Vallor 2024), but in many cases, they have introduced new harms that further exacerbate them (Chan et al. 2023). To correct those, the bulk of AI research focuses on correcting data, model, and outputs, i.e. focus on the technical system itself (called the *algorithmic frame*) (e.g. Mehrabi et al. 2021; Ferrer et al. 2021).

For instance, in the context of facial recognition systems that have been demonstrated to be disproportionately erroneous for Black women (Buolamwini and Gebru 2018), common actions within the algorithmic frame would include adding Black women’s faces to the training data or “debiasing” the algorithm to achieve equal error rates for Black women and other subgroups. However, situating the AI system in a social context and within power relations (what Selbst et al. (2019) refer to as *socio-technical frame*) would demonstrate that for example the dataset was trained under inhumane working conditions (Kong 2022; Miceli, Posada, and Yang 2022), or that the use of facial recognition for racial profiling in policing has been disproportionately harmful for black women (Eubanks 2018). Thus, *only* taking action within the algorithmic frame has been widely criticized for a limited effectiveness in advancing social justice (Kasirzadeh 2022; Bolte and van Wynsberghe 2025). In the specific context of intersectionality, this is expressed by an over-reliance on legally protected (sub-)groups that are detached from power (see e.g. Mirsch et al. (2026) for an overview, also referring to Kong (2022); Ovalle et al. (2023); Hampton (2021); Kalluri (2020)), not considering the societal impacts from AI outcomes or the mere choice of using an algorithm at all (Hoffmann 2019; Davis 2023; Bender and Hanna 2025).

2.2 Related Works

Using the Lens of Intersectionality to Expand the Algorithmic to the Socio-Technical Frame. Multiple recent studies suggest using an intersectional lens to extend AI practice toward a socio-technical frame (D’Ignazio and Klein 2020; Kong 2022; Ovalle et al. 2023; Mirsch et al. 2026; Vethman et al. 2025), some of which suggest concrete ways forward. Most notably, Vethman et al. (2025) move beyond abstract statements (too high-level guidance being a key barrier to action (Sadek et al. 2025)) and involve the target audience of AI practitioners to co-design concrete recommendations. Together with the recently published after Mirsch et al. (2026) which has a similar aim, these works provide a conceptual grounding for aligning AI practitioner’s practice with the Black feminist ideas of intersectionality beyond technical fixes. Their recommendations form the base for our work (details in §3.1). As such, this contribution extends previous work through direct representation of CSOs. This differentiation directly relates to an intersectional lens, as it puts front and center the knowledge of those directly affected by possible harms.

The Role of Civil Society Organizations. Indeed, there are many efforts, both with and without academic involvement, that have showcased the necessity and power of centering relevant subgroups of civil society, often under the umbrella term of “Participatory AI” (Bondi et al. 2021). For example, Allhutter et al. (2020) collaborated with CSOs to hold public bodies discriminating against job seekers accountable, and Gadiraju et al. (2023) centered the community facing harms from ableism to formulate what they require from large language models. Still, there is also increasing critique on shallow participatory AI practice (e.g. the exploitative nature of mere consultation, with little to no

influence to decisions, tokenism and cooptation) and various open challenges in centering civil society remain for it to be meaningful in the eyes of civil society (Birhane et al. 2022; Bondi et al. 2021; Katell et al. 2020; Dennler et al. 2023; Queerina et al. 2023; Erete, Rankin, and Thomas 2023). In line with these related works, we carefully considered the role of civil society within this research effort (to avoid blindly reproducing structural injustices embedded in scientific research practice), primarily through involving CSOs as equal partners (next to academic partners). We consciously center their experiences in working towards social justice *across* projects and organisations in our research methodology (see §3.2). This differs from previous works that centered CSOs for detailed reflections on specific applications, community participation, or organizational structure. Through our focus across contexts and communities, we distill more general CSO priorities from those emerging in specific applied experiences, and then transform these insights in reflections and suggestions for AI practitioners. Not to claim that through AI practitioners is the only or most important pathway to social justice, rather, it is to embolden *motivated* AI practitioners to align with and complement CSOs’ experienced practices in achieving success for social justice.

3 Methodological Approaches to Center Civil Society Voices

To gather CSO perspectives, we conducted a CSO community workshop (§3.2) and deep-dive interviews (§3.4), both analyzed through thematic analyses (§3.3 & §3.5).

3.1 Baseline Actionable Recommendations

We first introduce the actionable recommendations that serve as the baseline for our study. Using a set of recommendations or toolkits facilitates nuance and depth and this has proven promising in advocating for social justice work in companies (Pettersen et al. 2026). As established in §2, we follow the recent recommendations by Vethman et al. (2025) since they include concrete recommendations of prominent earlier social justice-oriented and intersectional works and directly involved the target audience of AI practitioners to ensure applicability. We expanded these recommendations through insights from Mirsch et al. (2026). Published shortly after, they proposed concrete desiderata to align AI practitioners’ practices with the socio-technical frame, focusing especially on intersectional Black feminist literature. While being largely consistent in the results, Mirsch et al. (2026) proposed three additional ways forward that we added to recommendations of Vethman et al. (2025). Figure 1 illustrates the final list of recommendations that were discussed during our workshops and interviews.

3.2 CSO Community Workshop

To gather critical CSO reflections on this set of recommendations (RQ1), we conducted an online workshop in November 2025 with 19 CSO participants, lasting 120 minutes. All participants held current or previous work experience at the intersection of data/AI and social justice, discrimination, or

Actionable Recommendations

- Collaborate with multiple disciplines before going into technical details.
- Use central role of AI experts to invite other disciplines and share responsibilities.
- Dedicate time and effort to create a psychologically safe environment.
- Write a positionality statement and reflect on it.
- Document perspectives and decisions throughout the lifecycle of AI.
- Discuss assumptions of data and AI.*
- Invite communities to co-own the participation process.
- Make participation financially sustainable for communities.
- Design a mechanism where impacted communities can safely voice concerns.
- Position the AI within social context and define the present power relations.
- Redefine concepts with historical and applied context.*
- Examine impact with complexity of intersecting social categories.*
- Be critical on your objective with data and metrics.
- Augment quantitative approaches with qualitative research and participatory design.
- Document clearly on the intended use and limitations of data, model and metrics.

Figure 1: Overview of the actionable recommendations from Vethman et al. (2025). Recommendations with an * are slightly altered based on Mirsch et al. (2026).

human rights. Ensuring **privacy** was of utmost importance. All participants consented to the publication of their contributions in a de-identifiable manner. Several voiced concerns that association with terms such as social justice or intersectionality could jeopardize important funding. Thus, we introduced the Chatham House rules, refrained from recording the meeting, and will disclose only minimal demographic details. Participants were recruited through public outreach via LinkedIn, through individual and through professional networks (Women in AI, Women At The Table and AI & Equality Hub). We made a conscious effort to invite potential participants with experience outside Europe and North America. Five participants have work experience in CSOs in Africa, two in Asia, fourteen in Europe, three in North America, and two in South America (multiple experiences per participant possible). To provide a *psychologically safe environment*, we explicitly laid out ground rules for psychological safety based on (Edmondson 1999, 2018). To address *agency and collaboration*, we designed the workshop so that participants would have an active speaking/writing role for most of the online meeting.

The **workshop design** comprised three parts. The first aimed to facilitate a shared understanding, activate collaboration, and set the scene. By inviting participants to recount success stories on defending human rights in the context of AI, we created the opportunity to network while emphasizing both the sense of urgency and celebration. In the second part, we presented a fictional, yet realistic high-stakes use case in the context of social justice issues within AI-supported fraud detection of social welfare. This use case demonstrates urgency with (financial) livelihood at stake and has documented adverse impacts around the globe (OECD 2025; Amnesty International 2024a, 2021, 2023). For the main part, participants collaboratively reflected on the actionable recommendations along the following questions: *Do you think the proposed actions are useful? Is something missing?* In sets of three recommendations, we connected the proposed actions to the use case, and highlighted the anticipated change. Participants gathered ideas on sticky notes on a whiteboard and held an extensive discussion afterwards. The workshop concluded with a dialogue on shared ideas

beyond the structured interactive exercises.

3.3 Thematic Analysis for RQ1

A thematic analysis (TA) is a qualitative research method to identify, analyze, and report on patterns in written text or speech which we applied to analyze the sticky notes as well as context given by our notes. We opted for the inductive reflexive TA (Braun and Clarke 2006, 2024, 2022), as it allows to structure the findings based on our interpretations of the input from CSOs rather than relying on predefined theoretical frameworks. The variant also considers the divergent backgrounds of the analyzers (us, authors) as an asset, as the method supports multiple interpretations per participant contribution. The analysis was supported by the software ATLAS.ti (v9.12.1), an interface and dashboard that simplifies coding and analysis, whilst we emphasize that all analysis was still performed manually.

We followed the six rigorous steps by Braun and Clarke (2024). (1) *Familiarizing ourselves with the data* entailed reading and discussing the workshop sticky notes. To (2) *generate initial codes*, we interpreted and paraphrased the core of the sticky notes, resulting in 57 unique codes for 65 sticky notes. A key categorization concerned whether a note was a critical reflection, a way forward (proposed action), or both. Following that, we grouped the codes into bundles to (3) *search for themes*. We focused here on the critical reflections and identified six key reflections as central themes, along with ten sub-themes. When (4) *reviewing the themes*, we slightly adjusted the theme on the costs of positionality after feedback from two workshop participants that we contacted to ensure correct interpretation. Subsequently, we (5) *defined and named the themes* and performed the final analysis by the (6) *act of writing* them in this paper.

3.4 Interviews

To dive deeper into proposed actions (RQ2), we conducted semi-structured interviews with CSO advocates or collaborators. We presented them the critical reflections that emerged from the workshop to reflect on promising paths forward. Our participants were experienced in (1) directly collaborating with or teaching (motivated) AI practitioners and (2) experienced on social justice, human rights, or inclusion. These selection criteria ensured that interviewees had the necessary understandings and shared language to move the critical CSO reflections towards concrete ways forward that were applicable and understandable to AI practitioners. As such, the interviews provide additional depth to our findings through longer nuanced discussions on a few concrete suggestions, based on their particular expertise and experience rooted in practitioners' contexts. We performed ten interviews between 40 and 120 minutes in November and December 2025 through video conferences in Google Meets. Their professional roles ranged from advisor, workshop facilitator, researcher, teacher, and ethnographer to activist in AI, Human Rights, Intersectionality, Governance, Psychological Safety, and diversity, equity, and inclusion. Six of the ten interviewees work or have worked in CSOs. Informed consent was given freely for the optional audio recording, which was only kept for two weeks to verify transcription,

deleted thereafter. After the analysis, the transcripts were kept pseudonymous without personal information until submission. Only the anonymous relevant snippets were kept until publication.

Interview design: Each interview began with a brief overview of the research context, recommended actions, and the findings of the CSO community workshop. Then, each interviewee was asked to choose which of the critical reflections they would like to discuss alongside the questions: *Do you recognize or have experiences related to the critical reflection or proposed action from the workshop?* and *What would be your proposed action for the critical reflection?* Within the time frame, each interviewee was able to reflect on three to five aspects. At the end of the interview, each participant was asked to provide a free-form reflection on ways forward, i.e. how to support the implementation of the proposed recommendations. After conducting five interviews, we observed some degree of repetition, but certain topics remained under-discussed. Therefore, we conducted five additional interviews. During the last two, we observed sufficient saturation, as we judged to have a substantial amount of suggestions per critical reflection, and only few comments were new relative to previously collected input.

3.5 Thematic Analysis for RQ2

The thematic analysis for RQ2 was conducted similarly to the first one (see §3.3), but on the proposed actions from the workshop and interviews. Interview transcripts were generated by the Google Meet application and checked and adjusted where needed. We again started by (1) *familiarizing ourselves with the data*; i.e. the interview transcripts and workshop sticky notes. (2) *Generating initial codes* for the interview transcripts resulted in 52 unique codes for 169 snippets. We then grouped the workshop and interview codes on ways forward into bundles to (3) *search for themes*. Multiple whiteboard discussions led to five central themes, each representing a direction for the motivated AI practitioner: the self, the team, the organization, the public, and the project manager. While (4) *reviewing the themes*, we decided to link the ways forward directly to the six key critical reflections derived in §3.3, since this aligns better with our aim to center the civil society voices. This resulted in 16 ways forward linked to the critical reflections. (5) *Defining and naming the themes* led to our final results, which are summarized in Fig 2. In the final iteration, each overarching theme name represents both the associated critical reflections and ways forward. We finished the analysis through the (6) *act of writing* this paper.

4 Critical Reflections and Ways Forward

This section presents the results from our thematic analyses, i.e. insights on critical reflections of CSOs on recent recommendations for AI practitioners designed to promote social justice (RQ1) as well as additional ways forward (RQ2). As illustrated in Figure 2, §4.1 to §4.6 will present critical reflections and ways forward jointly across our two data collection activities (workshop and interviews). Anonymous quotes from the workshop are tagged with [W] and those from the interviewees with [I1] to [I10].

4.1 Humility and Safety as Preconditions for Valuable Interdisciplinary Collaboration

Critical Reflection Workshop participants agreed that promoting interdisciplinary and safe collaboration is crucial, yet identified three challenges in practice. First, it might not be straightforward to “**identify which disciplines** or expertise should be invited”[W], since this is likely to be context-dependent. Second, even **interdisciplinary teams might be homogeneous demographically**, and still not include people from communities disproportionately harmed by AI. Participants highlighted the challenge to incentivize an interdisciplinary “team consisting of people from different communities”[W]. Third, even in interdisciplinary teams, there might be a “**hierarchical relationship between disciplines**”[W] where technical or quantitative efforts are valued more than social science or qualitative efforts.

Ways Forward Firstly, although the precise composition of an interdisciplinary team is likely to be contingent on contextual factors, the workshop participants suggested a potential establishment of a **minimum baseline**, informed by successful experiences. One interviewee stated that, based on their experiences with the public sector, minimum requirements would be: two technical experts, the project manager, a legal expert and at least one or two domain experts, e.g. “a person currently at the core of the process in which AI is to be introduced or with known experiences with citizens on the receiving end of the public service in question.”[I1] Interviewees emphasized that introducing additional disciplines may have a limit, and argued to only invite those willing to “actually spend the time and money to engage”[I5] to prevent “a chaotic conversation where not everyone is as involved.” [I1] Multiple interviewees shared the observation that setting overly high requirements may result in the complete disregard of interdisciplinarity.

Secondly, interviewees emphasized that, in addition to the interdisciplinarity itself, “it’s all about the quality of the conversation” [I8], meaning “possibly it matters who is in the room but it matters more how you are in the room together”[I8]. One interviewee asserted that “safe brave deliberation processes stand above the need for more and more voices”[I5], and referred to the concrete forms of Arao and Clemens (2023). Workshop participants had similarly voiced that in the absence of a **safe space for deliberation and reflection**, “the assessment can quickly become a ticking exercise.”[W] Especially for voicing resistance against the deployment of an AI system, as “people feel uncomfortable about the idea of refusing”[I5], a safe space for deliberation was considered non-negotiable.

Thirdly, to circumvent hierarchical relationships between different disciplines, one interviewee made reference to the concept of “**epistemological humility**”, which they defined as follows: “humbleness that leads to legitimization of everybody having a say”[I3]. This was positioned as vital to recognize the value of other (non-technical or non-quantitative) expertise, methodologies, and sources of knowledge. It was suggested that a period of reflection be allocated to each discipline in order to recount the value, limitations, and key assumptions of the proposed methodologies. This would not

THEMES	CRITICAL REFLECTIONS			WAYS FORWARD		
4.1 Humility and safety as preconditions for valuable interdisciplinary collaboration	Required disciplines depend on context → hard to identify	Despite diversity in disciplines, demographics remain homogenous	Hierarchy between disciplines: Technical over social sciences	Starting point: minimum baseline of disciplines	Safe spaces are key in attracting and keeping diverse teams	Stimulate epistemological humility through transparency about values, limitations, & assumptions
4.2 Evaluating AI's desirability to tackle a problem (vs. non-AI alternatives)	Formulating AI as potential solution at odds with omnipresent techno-solutionism	Lack of processes to compare AI to alternatives and to scrutinize the status quo		Problem-first approach: Examine AI's contributions to solving the problem (vs. technical objectives)	Desirability check: Adapting goals & timelines in project proposals to allow for success even when refraining from deploying AI	
4.3 Transparency as insufficient yet necessary condition for accountability	Transparency efforts are overstated and insufficient to enable accountability			Documentation beyond technical aspects, including who can be held accountable	Clear and accessible channels to give feedback & contest → enable deliberation	Dispelling AI myths is key to invite and enable critique
4.4 Meaningful participation requires long-term CSO partnerships & direct community engagement	Trust is a challenging precondition for participation	Involvement of CSOs cannot replace direct engagement with affected communities		Clearly communicating expectations, feedback, disagreements, & redress is key for trust	Building long-term and trust-based partnerships	Capacity building complements CSO involvement with direct community involvement
4.5 Extending positionality statements to foster action by disclosing costs around unequal burdens	Statements on positionality rarely incentivise action since they are distanced from associated costs			Highlight investments in diversity efforts through including incurred costs in positionality statements	Stress further investments towards overcoming exclusion barriers	Positionality statements should include positions of adversely affected groups
Meta-Reflection						
4.6 Responsibility of AI practitioners and organizational dynamics as obstacles & opportunities	Power differentials inside of organisations impact possible spheres and actions of influence	Technical practitioners often defer ethical responsibilities to colleagues		Revoke limited spheres of influence as excuse for inaction	Increase practitioner awareness through fostering critical competencies	Recognize layered spheres of influence: Prioritize strategic allyship and code of conduct

Figure 2: Key themes derived from CSO workshops and interviews, each with critical reflections and ways forward.

imply an expectation that all individuals become experts in all other disciplines; rather, it would enable the recognition of the fundamental principles of collaborators' expertise and foster effective communication. One interviewee concretized this in group projects with diverse roles where students realize that the "soft skills to communicate with someone who doesn't have the same background as you"[I5] are crucial to achieving success in their field.

4.2 Evaluating AI's Desirability to Tackle a Problem (vs. Non-AI Alternatives)

Critical Reflection Workshop participants agreed that AI systems can be a promising means of tackling a problem, but should never be a goal in themselves. This is "incredibly important and often lost in these evaluations and community engagement discussions"[W]. Seeing AI as a means, however, entails reflecting on *whether* to use AI systems rather than *how* to use them in the first place. This, in turn, might be **at odds with a techno-solutionist view**, where technology is seen as a viable solution to most problems. With incentives of profit, efficiency, AI hype, or the fear of falling behind, bringing the option to "not deploy at all" might be unpopular. One participant raised the doubt whether projects can be funded if they are positioned to question whether AI will be used rather than how. To examine *whether* the use of AI systems is desirable requires undertaking a comprehensive examination of the prevailing circumstances and to evaluate the viability of alternative solution strategies. For example, in the context of high-stake algorithms employed for welfare fraud detection or immigration services, this entails

a thorough interrogation of the advantages and disadvantages of the prevailing public service arrangements. Workshop participants expressed concerns whether they "have the capacity to critically reflect on how currently public services are run."[W] Consequently, a potential challenge may be the **lack of processes to compare AI to alternatives and to scrutinize the status quo**.

Ways Forward Two approaches were proposed to address the aforementioned challenges. Firstly, interviewees proposed initiating a reflection process to achieve awareness and an explicit definition of the AI system's objectives in their entirety, and to establish the meaning of success in terms of a problem-first approach. Consequently, a reformulation of success would entail **examining the AI system's specific contributions to resolving the prevailing problem, rather than a narrow focus on technical objectives** such as accuracy, efficiency, or productivity. One interviewee shared specific strategic questions that had helped in their collaborations with tech professionals and students. These questions were used to assess the need for a new product, reflect on potential harms, and determine whether one would be comfortable using the product while knowing how it works and whether one would recommend it to someone who does not know how it works. Such a strategy could be useful "to get technical developers to spell out what benefits they think they're bringing to the particular problem that they're trying to solve"[I4]. Questions could include: "1. Do products or solutions already exist for which we consider AI? If yes, are they good or bad solutions? (...) 2. Does our proposed solution potentially cause harm? And if so,

how are the harms relative to alternatives? (...) 3. Could and would you use it yourself? And would you recommend the product to someone else?”[I5].

Secondly, interviewees stressed the importance of “valuing non-acting as a successful output of one of those processes”[I3]. To successfully do so, a **desirability check** can be included into the project already at the proposal stage: rewriting the project as “a two-stage exploration, is one way to do it. Having a first phase to validate the project, to focus on desirability before even feasibility.”[I3] Participants stated that this preliminary period on desirability is ideal to “foster open communication between different departments and exchange information on the suitability of using AI”[W]. Ideally, this desirability is monitored in “regular compliance checks, regularly after introduction, keeping track of the system”[W]. Although costly at first glance, this reformulation of success with a desirability check might offer an alternative means of achieving project success. It could lead to the selection of an alternative solution to a problem and prevent spending on feasible yet undesirable AI projects, ultimately resulting in cost savings.

4.3 Transparency as Insufficient yet Necessary Condition for Accountability

Critical Reflection Workshop participants reflected on widely spread transparency efforts for documentation and public disclosure. On the one hand, these were considered essential and a basic right, especially for algorithms in public services: “Affected persons have the right to be informed that a system was used, an explanation for any decision and the ability to contest the decision”[W]. On the other hand, participants criticized their experience with overstated and insufficient efforts: publicly available documented efforts (e.g., on websites) for transparency were shallow and/or limited to technical documentation. *Why* an algorithm was used in the first place, is often left obscure. Many mechanisms for transparency, such as algorithm registers, would fall under the *trust but verify paradigm*, meaning that the burden of responsibility is shifted to journalists or CSOs to review the information and, potentially, hold parties using AI accountable. Thus, although **transparency efforts are important in themselves, they would not usually be sufficient to enable accountability**.

Ways Forward Firstly, the interviewees agreed that, based on their experience, efforts to increase transparency are often mistakenly considered sufficient for accountability purposes. Using the example of a national algorithm register, one interviewee explained: “it is clear that an algorithm is used. There is *some* level of information available about the why and who”[I1]. However, especially when part of the innovation is developed in-house, frequent hand-overs of responsibility between IT/AI development departments to executive departments tend to muddy overall accountability. Thus, workshop participants asked for more detailed public information on “**who is accountable and how they can be held accountable**”[W].

However, the interviewees emphasized that responsibility is not just about taking accountability when things go wrong. Instead, democratic deliberation is also a crucial part of ac-

countability. Transparency can therefore be an important means of fostering deliberation and public scrutiny, which may “ultimately lead to democratic governance”[I7]. This would require **clear accessible channels for feedback**, contestation, and redress. Interviewees particularly urged that these channels have *empowerment* as a guiding principle in their design to inspire collective action for social justice. For example, a “self-reporting mechanism where the users themselves could actually report issues”[I2] needs to go beyond having easy ways to report if some harm occurs and include “how many reports have been filed already. Then you can see you’re part of a bigger community of users that actually take the extra step and report harm”[I2].

As an important prerequisite for stepping into honest deliberation, interviewees addressed the opportunity for practitioners in their communication strategies, particularly within the light of **dispelling the AI myth**. Even though the tech industry is interested in powerful metaphors about AI fixing all societal problems, there are “plenty of figures saying that a lot of organizations are investing massively in AI and seeing little to no return on investment”[I3]. Interviewees described taking accountability as being clear about what AI is and is not, and what it can and cannot do. This was sketched as a matter of will, but also in terms of competence to prevent your message being “trapped in this echo chamber”[I2] of people already critical of AI and to make it reach investors, citizens, and users of these tools. Interviewees also urged for communication that challenges the inevitability of AI, as we “need to understand it, but you definitely shouldn’t evangelize it”[I5]. Proposed strategies were to elucidate the inner workings of AI (e.g. that chatbots do not have the human cognitive abilities to think or reason, as they are alluded to) and to introduce “counter-metaphors”[I2] for demystification, with ideas ranging from AI not being a magical floating brain or a cloud, “but rather made of data centers that consume the water in communities that already face droughts”[I2].

4.4 Meaningful Participation Requires Long-Term CSO Partnerships and Direct Community Engagement

Critical Reflection Workshop participants agreed on the importance of co-ownership, remuneration, and “long-term sustainable structures and safe and transparent decision pathways”[W] in the context of community engagement. Most workshop participants had extensive expertise in this and reported two key challenges: Firstly, they mentioned **trust as a challenging but required precondition for participation** and that effective trust-building activities are heavily context dependent. Secondly, workshop participants stressed that **CSO involvement cannot replace direct engagements with affected communities**. While CSOs can play a key role in participatory processes due to their expertise and close connection to affected communities, CSO representatives in the workshops were reserved in their own role being spokespersons of affected communities. “It’s important to open the space to as many communities as possible to ensure there’s full representation of lived experiences, as direct representation”[W].

Ways Forward Participants highlighted that humility and openness were paramount in order to build trustworthy collaborations. This requires very **clear communication strategies** as well as multiple channels through which people can (dis)agree, voice concerns, or seek redress: “it should be clear what happens if the safe process is rendered unsafe by a player. This may reduce the rejection and convey the commitment”[W]. Next to these channels, workshop participants considered it a substantial signal when (public) parties already plan to include the option (for citizens) to opt out of using a system, signaling that possible adverse impacts of AI systems were taken seriously. Meaningful participation was argued to necessitate **long-term “relationship-building with communities”**[W], especially “since situations change and communities should be allowed to change their mind”[W]. Inviting CSOs from the start with long-term intentions, stimulates asking “their motivation or goals beforehand”[I9] rather than just being asked for feedback when “established ideas have hit a wall”[I6].

Finally, interviewees agreed that collaboration with CSOs is still vital: “Taking some random people might not encounter empowered active people. Participants need to regard themselves as a relevant group for the study. When collaborating with or through a CSO then their specific interests are already reflected on”[I9]. On the other hand, they also stressed **complementing CSO involvement with direct engagement with people at the receiving end of AI**. This entails capacity building, i.e. empowerment so that communities in subsequent cases can contest without or with less support: “They didn’t know about this harm at all until we told them and we [pointed] them to literature and they started to reflect and host their own community assembly.”[I4]

4.5 Extending Positionality Statements to Foster Action by Including Costs on Unequal burden

Critical Reflection Our workshop participants reflected on the recommendation to explicitly declare the positions of the team developing the AI system, e.g. through so-called positionality statements. These are free-form reflections that situate the practitioners’ standing in the context of their work and make their perspectives, blindspots, and assumptions transparent. Although appreciated, participants noted that such statements rarely incentivize action: “Positionality alone is not enough, we need actions and investments to rebalance positionality”[W]. By *rebalancing positionality*, they referred to course-correcting measures that would integrate the perspectives of people within society who are affected by the AI system, thus bearing the costs if harm occurs: “It is not enough to recognize that disenfranchised people are not at the table, course-correcting investments are needed to alleviate these positionality imbalances”[W]. These harms, as well as course-correcting measures to prevent them, are costly; for example, to involve more people of a variety of disciplines, engage with communities, and take time to allow teams to reflect on their own position. **Since economic incentives are a common driver of action, workshop participants critically reflected on the lack of associated costs in current positionality statements to incentivize the associated action.**

Ways Forward To promote more active conversation around taken or required investments to improve the diversity of team perspectives, one participant proposed addressing the associated costs of positionality in a so-called “positionality cost statement”[W]. This resonated with other participants who suggested that such statement could contain the costs of the following three different interpretations: Firstly, the **costs of investments already made towards achieving a diversity of perspectives**: e.g. additional hires, psychological safety trainings, or expenses from successful/failed attempts for community engagement. Secondly, the participants highlighted that not only investments made but also **further investments required should be reported**, i.e. costs or estimates (e.g. for including people from disenfranchised communities) that are considered necessary but did not fit in the current budget. The third concerns extending the positionality considerations of the team to those affected by AI initiatives, i.e. to also **include which positions in society bear the cost (harms) of the AI in practice**.

Interviewees appreciated the plurality of these interpretations. They additionally proposed environmental costs to be included within the third interpretation, to map out who faces the “environmental impact of the whole life-cycle, from mining to energy use to maintain, as well as disposal”[I7]. Another interviewee added that “the emotional, health and psychological costs”[I10] could be included within the first interpretation. Although often unrecognized, they play a large role in bringing critical reflection to AI research and practice. “Being in an environment where the majority is comfortable in not pushing back, those folks deviating from that norm are punished. They are often people in black and brown bodies, and may leave the space entirely”[I10].

4.6 Meta-Reflection: The Responsibility of AI Practitioners and Organizational Dynamics

In addition to the *content* of the recommendations, participants reflected on the *process and approach* of bringing recommendations to motivated AI practitioners more broadly.

Critical Reflections Workshop participants agreed with the generally underlying assumption of “treating AI as part of existing power relations, not as a neutral apolitical tool”[W], making it necessary to examine the structural dimensions of who can decide, who can contest, and who is on the receiving end of AI interventions. They stated that “developers *should* assess the amount of power they have to influence the project they’re working on”[W], but acknowledged that **their actual influence might be limited due to “power differentials existing inside the organizations** that deploy systems”[W]. Often, this is used to **defer ethical responsibilities**: multiple interviewees shared that critical reflection on the adverse impacts of AI is currently rarely practiced by AI practitioners. Instead, they noted structural delegation to legal teams, management, or a designated AI ethics advisor, a practice they criticized as being “deeply problematic because we all have points of intervention where we can make ethical choices and we should make them and this shouldn’t just be outsourced to one person in charge”[I2].

Ways Forward Interviewees shared the workshop participants’ concern that the influence of AI practitioners may be limited by organizational structures, but saw no contradiction in **holding both organizations and AI practitioners accountable for taking efforts for social justice**. Interviewees underlined frequent observations around practitioners’ limited critical reflective skills and language, a likely barrier to taking action for social justice. They attributed this partly to their education—i.e. that engineers might not be confronted with the extent of their individual responsibility—and partly to the environment they encounter every day, where AI is seen as desirable, neutral, and apolitical. One interview conveyed this as an “extension of techno-solutionism that we label sometimes techno-exceptionalism”[I3], i.e. “that tech innovation is an exception to any forms of demands”[I3] which leads to the common view that “the people that pursue it are not accountable for the consequences of their technology”[I3]. To overcome this, they **encourage AI practitioners to engage in meta-reflection** to recognize how the education system and their work environment might have limited their ability to be critical and to actively avoid the outsourcing of such tasks described above. However, the organization’s (political) strategy does still play a crucial role: One interviewee recounted experiences of multiple professionals in STEM jobs that were structurally asked to avoid the social context of the innovation in question or even “policed in doing anything that deviated from the *social context agnostic* work, as there is an unspoken norm of decontextualizing”[I10].

For organizations motivated to move towards social justice, however, our participants suggested several practices with which **organizations can support AI practitioners**. “As the demand for AI innovation often is pressured from above”[I6], lasting change requires **support at the “higher management** who decided in the first place that the innovation was necessary”[W] and is ideally “mandated through corporate governance”[I8]. To facilitate interdisciplinary collaboration on the team-level, organizations can additionally provide a **moderator**: on the level of “the interdisciplinary team, we need someone, a moderator, to oversee how we do it together”[I8]. Such a moderator should “know the profession content-wise, and know the struggles, and able to pinpoint them”[I8]. Concretely, they can e.g. propose a code of conduct to foster individual responsibility in the employees, as a shared commitment “that they can and should refuse any tasks they believe are not in favor of the social well-being”[I8]. As “the link between the levels” [I9], understanding both management and bottom-up drives, the moderator can signal what to start with, “the low-hanging fruit that would work to build relationships and trust before you try to reconfigure their actual practices as opposed to just kind of making small changes at first”[I5]. In addition to these formal support structures, two interviewees emphasized the importance of having **allies to combat power** differentials in organizations: “An internal support system is essential, it may come in many forms, sort of like informal ethical working groups that are under the radar in which they can internally in a safe way discuss issues such as a manager dismissing ethical concerns”[I1].

5 Discussion

This study was designed to align recommendations towards social justice for AI practitioners with the priorities of those actively fighting for it: civil society organizations. CSOs supported existing recommendations, yet had clear priorities and necessary additions to bring practice closer to their experiences when working towards social justice (§4). We summarize key findings before outlining limitations, future work, and promising recommendations towards practice.

5.1 CSOs as Experts in Structural Issues

The insights from our CSO participants improve the initial recommendations by adding nuance and ways forward. This positions the CSO community as an extremely valuable, yet often overlooked source of experience and expertise regarding social justice-oriented AI practice. While our baseline recommendations (§3.1) were grounded in socio-technical and intersectionality-related literature as well as AI practitioner experiences, our CSO participants identified key considerations around structural challenges and injustices, for example, that techno-solutionism and the hierarchy between disciplines impedes mutual respect and learning (in line with e.g. Raji, Scheuerman, and Amironesei 2021; Bates et al. 2020). Whilst the contextual nature of socio-technical considerations are often emphasized, participants recognized each others’ structural challenges that occurred across their diverse projects. CSOs underlined a critical mindset, humility, and adaptivity (embodied by e.g. a desirability check, dispelling AI myths, or adding concrete costs to reflexivity) as vital for enacting change towards social justice—independently from the concrete use case.

5.2 Distinguishing Spheres of Influence: Start Small and Expand Outwards

Previous works have found that AI practitioners often refrain from taking actions towards social justice since they either fear that “being critical does not fit the work environment” or “to never do it good enough” (Vethman et al. 2025) (see also Sadek et al. 2025; Rakova et al. 2021; Schiff et al. 2020). Our findings address both concerns since our CSO participants explicitly acknowledged structural barriers and power imbalances within organizations that restrict possible spheres of influence or action (§4.6). In fact, they did not expect AI practitioners to completely overturn or contradict their organization’s operations but to *start within* their sphere of influence (§4.6). This reflects Young’s view of social justice, i.e. the notion that we all have a responsibility to endure some discomfort to recognize structural injustices and effect lasting change as we are all part of and benefit from the structures that make our society (Young 2011). In other words, we found that CSO experts express grace and understanding towards pragmatic and imperfect approaches, however, do not see structural challenges as an excuse for inaction. Indeed, our participants reported various ways forward that practitioners can immediately apply to their everyday work, e.g. to practice epistemological humility (§4.1), clearly communicating expectations, feedback, and values (§4.4), or to consider the value of AI (vs. other solutions) in

a problem-first approach (§4.2, (see also Bender and Hanna 2025; Katell et al. 2020; Gerards et al. 2022)). Finding momentum in a practitioner's current sphere of influence with the potential to slowly expand outwards. Indeed, we hope that the recommendations resulting from our study further this goal: while designed to provide a roadmap of social justice-oriented actions they may also help to communicate and legitimize such actions to colleagues (as in Petterson et al. 2026). This may in turn support inviting more AI practitioners to affect social change, i.e. to create organization-internal allyship (Ross 2025; Abebe et al. 2020).

It may appear paradoxical at first to advocate for *structural* changes yet targeting recommendations for motivated, *individual* practitioners. However, our findings emphasize that AI practitioners' choices are part of larger structures and have the potential to cause a ripple effect within their organization and beyond. Importantly, practitioners' recognition and action should be seen as a powerful complementary approach for social justice rather than a replacement for CSO initiatives (e.g. advocacy and litigation) tackling structural challenges that cross organizational boundaries.

5.3 Limitations & Future Work

This study has three limitations, each of which open up avenues for future research. First, we aimed to include diverse and underrepresented voices and achieved participation from at least two organizations from five continents for the workshop (§3.2). Still, dominant voices from Europe remained overrepresented and future work could explore how reflections differ when CSOs of other continents or regions are deliberately centered. Second, our findings might be affected by the specific use case the workshop focused on, i.e. fraud detection in public social welfare recipients. This was a deliberate methodological choice to gather insights anchored in a concrete case with documented cases of manifested harm around the globe (OECD 2025; Amnesty International 2024a, 2021, 2023). However, this anchoring may have skewed our findings towards automated decision-making algorithms in the public sector. To mitigate this, we explicitly raised this concern during the workshop, prompting participants to also share ways forward based on their *own* experiences, evoking responses that covered a broad range of high-stakes AI use cases. Nevertheless, we recommend future works to test the applicability of the recommendations to other domains. Third, during the workshop, CSO reflections were stimulated by existing, practitioner-oriented recommendations (§3.1) before moving to more open reflections. While we chose this format to provide nuance and depth to existing recommendations, this is likely to have guided reflections. We invite future works to apply these recommendations to specific local projects or take a more-open ended approach, which may confirm structural commonalities raised in our study or signal new aspects.

5.4 Recommendations Towards Practice

We now discuss how our findings can inform concrete actions towards social justice. First, we urge for actions that *increase the number of motivated AI practitioners* that are aware of our CSO-informed recommendations. Here, we see

a concrete role for practitioner-focused educators to integrate our recommendations into their *educational resources* such as those from AI & Equality Community (2023), Turing College (2025), The Alan Turing Institute (2023) or UNESCO (2026). Aligned with studies that advocate for reflective exercises (Bates et al. 2020; Boyd and Shilton 2021; Lindberg, Karlström, and Männikkö Barbutiu 2021), trainings for critical competencies (see §4.6) should also cover the value and methods of reflexivity.

Second, our findings show that the CSO community holds extremely valuable experiences and expertise on how those outside civil society can work towards social justice (§5.1). Thus, we urge to make this community, still too often overlooked in research contexts, a *fundamental partner in discussions around social justice-oriented AI development*. This could include future research efforts (see §5.3) as well as in concrete cases in public services, preferably where CSOs have already identified and addressed structural issues (building on §5.2) (see e.g. Coders Beyond Borders 2026). Additionally, while it is not new to invite CSO voices into policy discussions, our findings suggest that this could go beyond the typical role of "skeptics" (p.3 in Barakat 2024) critiquing suggestions. Instead, CSOs could be an integral part from the outset, i.e. co-author guidance on social justice-oriented AI in a more foundational role. This is essential given the dominant role of large technology corporations in the AI governance discourse (Birhane et al. 2026) as well as gaps in socio-technical understandings in policy perspectives (Migineishvili et al. 2026).

Third, towards putting these recommendations into practice and following the "spheres of influence" approach (§4.6), it seems promising to *look for application areas with incentive to focus on social justice*. Here, the public sector seems especially suitable: it is (i) under obligations to work in line with public interests (Cobbe et al. 2020) (vs. commercial interests alone) and (ii) increasingly adopting AI in its services (Valle-Cruz, Gil-Garcia, and Sandoval-Almazan 2024) (key reasons for our case study, §3.2). Here, AI often disadvantages the same communities that are the most reliant on public services, making it crucial to consider AI deployment with great care and focus on social justice (Hudig, Kallina, and Singh 2026). Driven by these motivations, the public sector could become an early adopter of our CSO-informed recommendations, e.g. by integrating them into processes of in-house development or making them procurement requirements (Hudig, Kallina, and Singh 2026).

6 Conclusion

The civil society actors, who have advocated and litigated against AI harms that exacerbate structural injustices, can also be fundamental allies to AI practitioners that are motivated to prevent the harms from happening in the first place. In this work, civil society insights sharpened recently formulated recommendations for practitioners into pragmatic actions that acknowledge, yet are not impeded by, structural and organizational barriers. We urgently invite practitioners to join us to complement civil society in fostering social justice through humble and strategic actions everyday.

Positionality Statement

As in the central aim of our contribution, we are grateful to reflect which positions are represented in the research design and writing of this paper, which voices are missed and which actions were taken to face our blind spots.

This research effort is conducted by a team based and with most experiences in Europe. As such, we actively faced our blind spots toward non-European experiences through discussions with colleagues in other continents and deliberate effort to invite participants from other regions in the world. Next to that, the team consists of a 50%-50% collaboration between critical academic actors and actors from civil society organizations. Both civil society partners, Women in AI and Women At The Table, are working on a global scale, with chapters and initiatives specifically targeting diverse non-Western perspectives. This position made us aware of how research can be extractive and exhausting for civil society. As such, we adjusted for example the format of the workshop (see §3.2) and proposed to give dedicated time back to participants for their initiatives, which some actors have acted upon.

Aware that not every relevant experience of oppression or compounded forms of oppression could be included in the team, we practiced what is eventually preached in this paper and in efforts before (i.e. Vethman et al. (2025) and Mirsch et al. (2026)): a humble, adaptive stance that aims to respect diverse opinions. We consciously reflected on people's diverse and unique realities and respected and welcomed any disagreement, discomfort or critical feedback between and within participants and authors. To illustrate this: during the workshop, a participant was critical of the notion that CSOs were considered as suitable representatives of harmed communities in the use case of welfare fraud detection. The participant wrote this on a sticky note and chose not to speak up when the author/facilitator invited them to elaborate. We respected this and instead asked all participants to share their arguments or experiences about civil society organizations being an unsuitable representative. That is, we acknowledged their voice, respected their reasons for not speaking up, and shared the burden of reflecting on this question with all participants to gather their nuance within the discussion.

We are aware of our position as actively learning actors within this domain of social justice and technology. Our chosen way of working, guided by the calling-in approach of Ross (2025), drives us to be very grateful for all the feedback received before, during, and after the research activities and makes us also welcome any constructive feedback or comments that interested readers of this article would like to share. As our commitment and initiative do not end with this research article, we are happy to take any thoughts with us in the continued effort to come.

Generative AI Usage Statement

No use of generative AI is used with respect to the preparing, generating or proofreading of this manuscript.

Acknowledgments

First of all, this study is partially funded by the DIVERSI-FAIR project from the Erasmus+ grant (101107969), and is therefore co-funded by the European Union. However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Next to that, this study was conducted within work of the Responsible Research and Innovation (RRI) Hub at RWTH Aachen University, as such partially funded by the Federal Ministry of Education and Research (BMBF) and the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) under the Excellence Strategy of the Federal Government and the Länder.

Finally, we are immensely grateful to the community members that invested time in discussion, mentorship and reviews: the anonymous AIES reviewers, Sacha Alanoca, Caitlin Kraft-Buchman, Teodora Groza, Alexi Orchard, Wilhelmina Ndapewa Onyothi Nekoto, Pierre Noro, Roxane de Pelet, Levent Sagun, Can Simsek, Faye-Marie Vassel, Carla Maria Verwer, Yung-Hsuan Wu, Raphaële Xenidis. And Ivy.

References

- Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Robinson, D. G. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 252–260. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- AI & Equality Community. 2023. < AI & Equality > Online Course. <https://aiequalitytoolbox.com/online-course/>. Accessed: 2025-11-17.
- AlgorithmWatch and Bertelsmann Stiftung. 2020. Automating Society Report 2020: Taking Stock of Automated Decision-Making in the EU. <https://automatingsociety.algorithmwatch.org/>. Accessed: 2025-01-10.
- Allhutter, D.; Cech, F.; Fischer, F.; Grill, G.; and Mager, A. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, 3.
- Amnesty International. 2021. Xenophobic Machines: Discrimination through Unregulated Use of Algorithms in the Dutch Welfare System. <https://www.amnesty.org/en/documents/eur35/4686/2021/en/>. Accessed: 2025-01-10.
- Amnesty International. 2023. Serbia: Trapped by Automation: Poverty and Discrimination in Serbia's Welfare State. <https://www.amnesty.org/en/documents/eur70/7443/2023/en/>. Accessed: 2026-01-09.
- Amnesty International. 2024a. Entity Resolution in India's Welfare Digitalization. <https://www.amnesty.org/en/latest/research/2024/04/entity-resolution-in-indias-welfare-digitalization/>.

- Amnesty International. 2024b. France: Discriminatory algorithm used by the social security agency must be stopped. <https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/>. Accessed: 2025-05-04.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Accessed: 2026-01-05.
- Arao, B.; and Clemens, K. 2023. From safe spaces to brave spaces: A new way to frame dialogue around diversity and social justice. In *The art of effective facilitation*, 135–150. Routledge.
- Barakat, H. 2024. Selective perspectives: A content analysis of The New York Times' reporting on artificial intelligence. The Maybe, <https://www.themaybe.org/research/selective-perspectives>. Accessed: 2026-08-05.
- Bates, J.; Cameron, D.; Checco, A.; Clough, P.; Hopfgartner, F.; Mazumdar, S.; Sbaffi, L.; Stordy, P.; and De La Vega De León, A. 2020. Integrating FATE/critical data studies into data science curricula: where are we going and how do we get there? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 425–435. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.
- Bender, E.; and Hanna, A. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. HarperCollins. ISBN 9780063418554.
- Berger, M. T.; and Guidroz, K., eds. 2009. *The Intersectional Approach: Transforming the Academy through Race, Class, and Gender*. Chapel Hill: University of North Carolina Press.
- Birhane, A.; Angius, R.; Agnew, W.; Pandit, H. J.; Mitra, B.; Dobbe, R.; and Talat, Z. 2026. Big AI's Regulatory Capture: Mapping Industry Interference and Government Complicity. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 2586–2607. New York, NY, USA: Association for Computing Machinery. ISBN 9798400725968.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394772.
- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. IEEE.
- Bolte, L.; and van Wynsberghe, A. 2025. Sustainable AI and the third wave of AI ethics: a structural turn. *AI and Ethics*, 5(2): 1733–1742.
- Bondi, E.; Xu, L.; Acosta-Navas, D.; and Killian, J. A. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436. Virtual Event USA: ACM. ISBN 978-1-4503-8473-5.
- Booth, R. 2024. DWP algorithm wrongly flags 200,000 people for possible fraud and error. <https://www.theguardian.com/society/article/2024/jun/23/dwp-algorithm-wrongly-flags-200000-people-possible-fraud-error>. Accessed : 2025-12-17.
- Boyd, K. L.; and Shilton, K. 2021. Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams. *Proc. ACM Hum.-Comput. Interact.*, 5(GROUP).
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.
- Braun, V.; and Clarke, V. 2022. *Thematic analysis: A practical guide*. Los Angeles, USA: Sage Publications.
- Braun, V.; and Clarke, V. 2024. Thematic analysis. In *Encyclopedia of quality of life and well-being research*, 7187–7193. Springer.
- Browne, J.; Cave, S.; Drage, E.; and McInerney, K., eds. 2023. *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford University Press. ISBN 9780192889898.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 77–91. Proceedings of Machine Learning Research (PMLR).
- Campbell, T.; and Mancilla, A., eds. 2012. *Theories of Justice*. Ashgate.
- Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krashennikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; Lin, M.; Mayhew, A.; Collins, K.; Molamohammadi, M.; Burden, J.; Zhao, W.; Rismani, S.; Voudouris, K.; Bhatt, U.; Weller, A.; Krueger, D.; and Maharaj, T. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 651–666. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Cobbe, J.; Lee, M. S. A.; Janssen, H.; and Singh, J. 2020. Centering the law in the digital state. *Computer*, 53(10): 47–58.
- Coders Beyond Borders. 2026. Our Work: Factstream AI. <https://codersbeyondborders.org/our-work/factstream-ai/>. Accessed : 2026-08-06.
- Collins, P. H.; and Bilge, S. 2020. *Intersectionality*. Cambridge, UK: Polity Press. ISBN 9781509539673.
- Combahee River Collective. 1978. A Black Feminist Statement. In Eisenstein, Z. R., ed., *Capitalist Patriarchy and the Case for Socialist Feminism*, 210–218. New York: Monthly Review Press.
- Council of Europe. 2024. Methodology for the Risk and Impact Assessment of Artificial Intelligence Systems from the Point of View of Human Rights, Democracy and the Rule of Law (HUDERIA). <https://rm.coe.int/cai-2024-16rev2-methodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f>. Accessed: 2025-01-22.

- Crenshaw, K. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics. *University of Chicago Legal Forum*, 1989: 139–167.
- Crenshaw, K. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6): 1241–1299.
- Davis, J. L. 2023. ‘Affordances’ for Machine Learning. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 324–332. Chicago IL USA: ACM. ISBN 9798400701924.
- Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; Talat, Z.; Russo, M.; and Pinhal, J. D. J. D. P. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 375–386. Montréal QC Canada: ACM. ISBN 9798400702310.
- D’Ignazio, C.; and Klein, L. F. 2020. *Data Feminism*. The MIT Press. ISBN 9780262358521.
- Edmondson, A. C. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2): 350–383.
- Edmondson, A. C. 2018. *The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth*. Hoboken, NJ: John Wiley & Sons. ISBN 9781119477242.
- Erete, S.; Rankin, Y.; and Thomas, J. 2023. A Method to the Madness: Applying an Intersectional Analysis of Structural Oppression and Power in HCI and Design. *ACM Transactions on Computer-Human Interaction*, 30(2): 1–45.
- Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press. ISBN 9781250112617.
- European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>. Accessed: 2025-01-13.
- Ferrer, X.; Nuenen, T. v.; Such, J. M.; Coté, M.; and Criado, N. 2021. Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, 40(2): 72–80.
- Gadiraju, V.; Kane, S.; Dev, S.; Taylor, A.; Wang, D.; Denton, E.; and Brewer, R. 2023. “I wouldn’t say offensive but...”: Disability-Centered Perspectives on Large Language Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 205–216. Chicago IL USA: ACM. ISBN 9798400701924.
- Gerards, J.; Schäfer, M. T.; Muis, I.; and Vankan, A. 2022. Fundamental Rights and Algorithms Impact Assessment (FRAIA). <https://www.government.nl/documents/2021/07/31/impact-assessment-fundamental-rights-and-algorithms>. Accessed: 2026-08-06.
- Hampton, L. M. 2021. Black Feminist Musings on Algorithmic Oppression. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, 1. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.
- Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.
- Hudig, A. I.; Kallina, E.; and Singh, J. 2026. “It’s Just a Wild, Wild West”: Harnessing Public Procurement as an AI Governance Mechanism. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–22.
- Kalluri, P. 2020. Don’t ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815): 169–169.
- Kasirzadeh, A. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’22*, 349–356. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.
- Katell, M.; Young, M.; Dailey, D.; Herman, B.; Guetler, V.; Tam, A.; Bintz, C.; Raz, D.; and Krafft, P. M. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.
- Kong, Y. 2022. Are “Intersectionally Fair” AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, 485–494. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Lindberg, S.; Karlström, P.; and Männikkö Barbutiu, S. 2021. Design Ethics in Practice - Points of Departure. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).
- McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference*.
- Meaker, M. 2024. Algorithms Policed Welfare Systems For Years. Now They’re Under Fire for Bias. <https://www.wired.com/story/algorithms-policed-welfare-systems-for-years-now-theyre-under-fire-for-bias/>. Accessed: 2025-05-04.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).
- Miceli, M.; Posada, J.; and Yang, T. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): Article 34.
- Migineishvili, N.; Gao, A.; Adjabodjou, A.; Thakur, D.; Just, R.; and Reinecke, K. 2026. Regulating AI: Where U.S. State Policy and HCI (Mis)align. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*,

- CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Mirsch, M.; Wegner, L.; Strube, J.; and Leicht-Scholten, C. 2026. Toward substantive intersectional algorithmic fairness: desiderata for a feminist approach. *AI and Ethics*, 6(1): 90.
- Monett, D.; and Lewis, C. W. 2017. Getting clarity by defining artificial intelligence—A survey. In *3rd conference on "philosophy and theory of artificial intelligence"*, 212–214. Springer.
- Naudts, L. 2024. The Digital Faces of Oppression and Domination: A Relational and Egalitarian Perspective on the Data-driven Society and its Regulation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 701–712. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.
- OECD. 2025. AI and the Future of Social Protection in OECD Countries. https://www.oecd.org/en/publications/ai-and-the-future-of-social-protection-in-oecd-countries_7b245f7e-en.html. Accessed: 2026-08-06.
- Ovalle, A.; Subramonian, A.; Gautam, V.; Gee, G.; and Chang, K.-W. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 496–511. Montréal QC Canada: ACM. ISBN 9798400702310.
- Petterson, A.; Kim Ly, C.; Cross, T.; Y. Wong, R.; and Chandra, P. 2026. Tooling Justice: Articulating Equity Work Through Design Toolkits. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Queerina, O. O.; Ovalle, A.; Subramonian, A.; Singh, A.; Voelcker, C.; Sutherland, D. J.; Locatelli, D.; Breznik, E.; Klubicka, F.; Yuan, H.; J. H.; Zhang, H.; Shriram, J.; Lehman, K.; Soldaini, L.; Sap, M.; Deisenroth, M. P.; Pacheco, M. L.; Ryskina, M.; Mundt, M.; Agarwal, M.; Mclean, N.; Xu, P.; Pranav, A.; Korpan, R.; Ray, R.; Mathew, S.; Arora, S.; John, S.; Anand, T.; Agrawal, V.; Agnew, W.; Long, Y.; Wang, Z. J.; Talat, Z.; Ghosh, A.; Dennler, N.; Noseworthy, M.; Jha, S.; Baylor, E.; Joshi, A.; Bilenko, N. Y.; Mcnamara, A.; Gontijo-Lopes, R.; Markham, A.; Dong, E.; Kay, J.; Saraswat, M.; Vytla, N.; and Stark, L. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1882–1895. Chicago IL USA: ACM. ISBN 9798400701924.
- Raji, I. D.; Scheuerman, M. K.; and Amironesei, R. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 515–525. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.
- Rawls, J. 1971. *A Theory of Justice: Original Edition*. Harvard University Press. ISBN 9780674880108.
- Ross, L. J. 2025. *Calling In: How to Start Making Change with Those You'd Rather Cancel*. New York, NY: Simon & Schuster.
- Sadek, M.; Kallina, E.; Bohné, T.; Mougnot, C.; Calvo, R. A.; and Cave, S. 2025. Challenges of responsible AI in practice: scoping review and recommended actions. *AI & society*, 40(1): 199–215.
- Schiff, D.; Rakova, B.; Ayes, A.; Fanti, A.; and Lennon, M. 2020. Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707*.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. ISBN 978-1-4503-6125-5.
- Sen, A. 2008. The idea of justice. *Journal of human development*, 9(3): 331–342.
- The Alan Turing Institute. 2023. Assessing and mitigating bias and discrimination in AI: Beyond binary classification. <https://www.turing.ac.uk/courses/assessing-and-mitigating-bias-and-discrimination-ai-beyond-binary-classification>. Accessed: 2025-11-17.
- Turing College. 2025. Certified AI Ethics course. <https://www.turingcollege.com/ai-ethics>. Accessed: 2025-11-17.
- UNESCO. 2026. Empowering Minds, Transforming Lives: Global MOOC on the Ethics of AI. <https://www.unesco.org/en/articles/empowering-minds-transforming-lives-global-mooc-ethics-ai>. Accessed: 2026-08-05.
- Valle-Cruz, D.; Gil-Garcia, J. R.; and Sandoval-Almazan, R. 2024. Artificial intelligence algorithms and applications in the public sector: A systematic literature review based on the PRISMA approach. *Research Handbook on public management and artificial intelligence*, 8–26.
- Vallor, S. 2024. *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.
- Vethman, S.; Smit, Q. T. S.; van Liebergen, N. M.; and Veenman, C. J. 2025. Fairness Beyond the Algorithmic Frame: Actionable Recommendations for an Intersectional Approach. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 3276–3290. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Xenidis, R. 2022. Intersectionality from Critique to Practice: Towards An Intersectional Discrimination Test in the Context of 'Neutral Dress Codes'. *European Equality Law Review*, 2.
- Young, I. M. 2011. *Responsibility for justice*. Oxford University Press.