AI & SOCIETY
https://doi.org/10.1007/s00146-025-02653-8

RESEARCH

Negotiating AI fairness: a call for rebalancing power relations

Marta Gibin[1] · Emmanouil Krasanakis[2] · Kęstutis Grumodas[3]

Received: 27 October 2024 / Accepted: 23 September 2025

Abstract

AI fairness is at the center of many debates, but there are different perspectives on what it entails. Currently, purely technical algorithmic fairness approaches dominate the scene, often neglecting a sufficiently well-rounded view of social implications and ignoring the voices of lay people. With the end goal of overcoming these issues, we investigate and synthesize the points of contact and differences among computer science, sociological and lay people perspectives and move towards a *lay-socio-technical* view of AI fairness. To this end, we conducted interviews with experts from the computer and social sciences, as well as a survey and co-creation workshops with lay people. Our results show that there are sometimes conflicting views on AI fairness across these perspectives. An integrated view requires two processes of negotiation: (a) between computer and social sciences, and (b) between experts from both disciplines and lay people. We identify strategies for supporting these processes, but we state that they are ultimately possible only if there is a rebalance of current power relations between the disciplines, and between experts and lay people.

Keywords AI fairness · Algorithmic fairness · Negotiation · Power relations · Interdisciplinarity

1 Introduction

The numerous cases of discrimination linked to the use of artificial intelligence (AI) systems (O'Neil 2016; Buolamwini and Gebru 2018; Obermeyer et al. 2019; Angwin et al. 2016) have stimulated reflections on the subject of fairness, both at academic and non-academic level. In academia, the topic of AI fairness is being addressed by various disciplines, each viewing the phenomenon through its own lenses and bringing its own perspective to the topic. On the non-academic side, concerns have been voiced over the use of AI systems in everyday life—from the spheres of healthcare and education, to transportation and entertainment, and many others. These concerns often culminate to whether system outcomes, such as decisions, predictions, or recommendations, are fair. For example, in the simplest cases, lack of fairness may be found in spurious connection between system outcomes and certain demographics, biased data used as AI inputs (e.g., some systems learn from racist jurisdictional decisions), or favor towards demographic majorities.

Although debates over AI fairness have grown in maturity, they have also given rise to a number of different—and often conflicting—perspectives among disciplines (computer science, philosophy, social sciences, law, etc.) and among stakeholders[1] involved in the process of defining what fair AI looks like. In this paper, we focus on three actors: computer scientists, lay people, and sociologists. To begin with, *computer scientists* are those implementing AI systems and ultimately responsible for the appropriate integration of fairness concerns, for example by following trustworthy development principles and measuring and mitigating specified types of

✉ Marta Gibin
marta.gibin2@unibo.it

Emmanouil Krasanakis
maniospas@iti.gr

Kęstutis Grumodas
kestutis.grumodas@protonmail.com

[1] University of Bologna, Bologna, Italy

[2] Centre for Research and Technology Hellas, Thessaloniki, Greece

[3] Vilnius University, Vilnius, Lithuania

[1] With stakeholders we refer to individuals or social groups who might be positively or negatively affected by the use of an AI system. They include, for example, those who develop the system, its users, the organizations profiting from it, policymakers, and vulnerable groups who might be discriminated against by its use. They may also include product owners, such as parent or funding organizations, that drive the system's main technical specifications in terms of functional requirements (what it does) and non-functional ones (its general properties).

Published online: 09 October 2025

bias. To address fairness concerns, computer science has created the emerging field of *algorithmic fairness*, an area of research focusing on strategies to ensure AI systems do not discriminate against individuals or groups. However, as of writing, the latter often relies on simplifications that make it hard to process the full range of real-world complexity and often lead to ignoring some valid but less frequently encountered fairness concerns. In order to address this issue, *lay people* are increasingly involved in AI design (Delgado et al. 2023) and asked to provide their view on what fairness is, as for example happened for the creation of the AI Act.[2] This aligns with goals for democratization, which is steadily emerging worldwide. Finally, *sociologists* study how these technologies are produced and the social consequences of using these systems, highlighting their potential risks and negative impacts. They can also serve as mediators between other actors, for example by providing support in gathering and organizing feedback from lay people.

In this paper we investigate the perspectives of these three actors on AI fairness and compare them with the current status of algorithmic fairness. We then try to meld these discussions into a *lay-socio-technical perspective* on AI fairness. We start from the idea that fairness is a perceived category that is negotiated among different actors. In the paper, we focus on two processes of negotiation: (a) how to combine the computer science perspective with the sociological one and (b) how to integrate the construction of AI systems with the opinions of lay people. These processes require negotiation because they involve forms of potential conflict and mediation. For example, if we consider AI systems used to decide on loan requests, the bank and the clients might have a different idea on what a fair outcome of the process would be. As further explored in Sect. 4.1.1, the bank might want the system to be impartial, while the clients might want their personal circumstances to be taken into consideration.

This manuscript reflects results from the work of the authors in the MAMMOth project,[3] an EU project that aims to build a toolkit that assists AI creators in the detection and mitigation of biases with a lay-socio-technical perspective. In line with this goal, we interviewed both technical experts and sociologists to further explore how they approach and intend the study of AI fairness. In parallel, we involved groups that, according to the literature (O'Neil 2016; Buolamwini and Gebru 2018; Eubanks 2018), are often at risk of being discriminated against by AI systems (migrants, LGBTQ + communities, women, etc.). These groups were involved through a survey and a number of co-creation workshops to further explore their ideas on what a fair use of AI systems would entail.

The article is structured in six sections. After this introduction, in Sect. 2, we briefly analyze the technical, sociological and lay perspectives on AI fairness. As these diverse angles offer different understanding of the same phenomenon, we speak of fairness*es* (in the plural). Sections 3 and 4 illustrate the methodology and results, respectively. In Sect. 5, we discuss the main evidence and compare the opinions on how to build fair AI systems from the three perspectives—and with the current state of algorithmic fairness. We then suggest some directions on how to promote a lay-socio-technical perspective on the topic. We conclude by emphasizing what we consider to be an overarching theme: the need to redefine power relations between disciplines and between experts and lay people, ensuring the opportunity for different views to find space in the discussion on how to build AI systems that are fairer.

2 Unpacking AI fairness(es)

In this article we focus on three approaches towards AI fairness: (a) the computer science perspective, also known as algorithmic fairness, (b) the sociological approach, and (c) the lay people approach. The technical understanding of algorithmic fairness mainly focuses on mathematical methods and numerical analysis. Although this approach has many benefits and is widely used in the industry, other disciplines—like sociology—argue for a broader, context-focused understanding of fairness. Such understanding takes into account the imbalances in power dynamics and puts an emphasis on not worsening the already existing inequalities. For this reason, sociologists call for including a diverse array of stakeholders in AI system's design processes and encourage a more bottom-up approach. Similarly, a lay understanding of fairness includes protecting personal data, tackling fear of discrimination and minimizing similar types of personal wrongdoings. Echoes of these viewpoints have been picked up by computer scientists, but have yet to become a cornerstone of algorithmic fairness research and development.

2.1 Algorithmic fairness

Algorithmic fairness research papers mainly follow four directions: (a) expressing a principle mathematically, and introducing algorithms to satisfy it in a specific scenario (Mitchell et al. 2021). Such works can serve as technical reference, but rarely recognize the negotiation that took place behind the scenes. For example, studies of real-world

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance.

[3] https://mammoth-ai.eu/ (Accessed 17/10/2024).

approval like those of Saxena et al. (2019), Bankins et al. (2022), and Starke et al. (2022) are rarely conducted. (b) Porting existing fairness assessment methods and algorithms to new data. For example, in addition to tabular data, fairness is also analyzed for images (Parraga et al. 2023), graphs (Dong et al. 2023), and large language models like ChatGPT (Zhang et al. 2023). (c) Improving existing algorithms in terms of computational efficiency, accuracy-fairness tradeoffs, explainability (Dodge et al. 2019), or the participation of humans in the decision-making process (Nakao et al. 2022). (d) Creating new fairness paradigms, ideally based on legal or philosophical definitions, which grapple with the discrepancy between abstract philosophy and unambiguous technical phrasing.

Mathematical definitions of fairness typically create predictive equilibria between individuals or population subgroups, and recognize imbalances of predictive measures (e.g., different misclassification rates) as a form of bias. There are three main technical schools of thought (Ntoutsi et al. 2020; Barocas et al. 2023; Castelnovo et al. 2021); *individual fairness* that enforces a maximal predictive differences between similar individuals (Dwork et al. 2012), *group fairness* that conditions predictive measure analysis on each population subgroup that contains overlapping sensitive characteristics (e.g., gender, ethnicity) (Calders and Verwer 2010), and *counterfactual fairness* that simulates real-world mechanisms with causal models and then removes bias from their outcomes through statistical methods (Kusner et al. 2017). Some definitions are contradictory in that, mathematically, they cannot be concurrently satisfied. They may also be at odds with the perceived business utility of AI (e.g., with its accuracy on selected datasets) in which case one may maximize utility subject to minimum fairness constraints, maximize measures of fairness subject to a desired level of utility, or optimize numerical trade-offs between fairness and utility.

Few approaches attempt to bridge the gap with social issues encountered in practice, as computer scientists are often not well-educated on the social ramifications their systems could create, and due to the technical (e.g., mathematical, implementation, system design) challenges of satisfying even conceptually simple definitions of fairness. In fact, algorithmic fairness often focuses on addressing simple challenges in practical frameworks (Bellamy et al. 2019; Bird et al. 2020), with the implicit assumption that practical adoption should be overseen by corresponding experts. However, this rarely happens in practice, which has led to attempts at mapping real-world considerations to technical definitions (Khan et al. 2021). This principlist approach (John-Mathews et al. 2022), albeit valuable in many contexts where fairness concerns can be modeled with existing mathematical formulas, encounters a barrier when ported to unforeseen settings.

In general, algorithmic fairness is one of the interplaying facets of trustworthy AI, which may also include other scientific and social science concerns. For example, it may not only be sufficient to make algorithms fair, but also make them robustly fair, where robustness refers to systems that are reliable in situations that are not encountered during their creation (Marcus 2020). One form of robustness is for small perturbations of inputs to induce only small changes to predictions. Aside from purely algorithmic practices, technical fairness may also include ongoing system monitoring, for example through processes known as TrustAIOps (Li et al. 2023) that incorporate ongoing user feedback about unfairness. One emerging area of research that serves as an example of how technical fairness research can be applied in the real world is that of algorithmic recourse (Karimi et al. 2021). This is a type of counterfactual fairness that not only provides explanations about bias or unfairness, but also suggests a series of human actions that may lead to desirable outcomes for individuals or fairer social outcomes.

2.2 AI fairness in sociology

Although it is the technical understanding that has dominated the subject literature, sociologists note that fairness must encompass much more than the technical side. In sociology, AI systems are analyzed as sociotechnical systems, where the social and the technical aspects are deeply intertwined (Sartori and Theodorou 2022). In particular, fairness is a complex concept that encompasses social, legal, ethical, and technical aspects. Therefore, reducing fairness to a mere problem of mathematical optimization risks ignoring the social context and the elements that cannot be fully computed and operationalized.

Sociologists note that AI systems often reflect the material basis of the social world in which they were built. This means that AI systems are deeply connected to power relations and this is clear from the many forms of discrimination that they perpetuate. The literature on the topic (e.g., Noble 2018; Crawford 2021) shows how power relations embedded in algorithms can perpetuate social inequalities and bias and explores the socio-political and environmental power dynamics involved in building and deploying AI systems. Systems built under white supremacy, capitalism, heteropatriarchy and colonialism should be examined as to not reinforce these same power dynamics and not to "perpetuate and amplify" the same power imbalances that already exist in society (Weinberg 2022). This can mean, for instance, fixating on representation of various marginalized groups within the data while ignoring the socio-economic conditions that produce those inequalities in the first place. The issue of inequalities has always been at the center of sociological research, and the extensive literature on the subject cannot be ignored when discussing the topic of fairness. Thus, sociologists emphasize the need for contextual understanding when it comes to AI fairness (Zhou et al. 2022).

The above considerations should be taken into account not only during the design of technical tools, but also in areas like the education of computer scientists (Luchs et al. 2023). In such areas there is usually a focus on collecting fair and representative data, but other areas should be considered as well, such as continuous evaluation of the system's fairness at all stages of development, honestly acknowledging the weaknesses of the system (Fenu et al. 2022). The creators of AI systems must embrace and understand their active role in creating such systems, thus focusing on the process of constructing meaning from data (Barabas et al. 2020). Due to the risks of reproducing existing unequal power dynamics and over-fixating on the technical, it is important to consider whether the systems should be built in the first place (Weinberg 2022) and avoid techno-solutionism, which refers to the idea that technical solutions are always the best way to solve complex social problems (Morozov 2013).

Sociology examines whose values are embedded in the systems and the power relations they (re)produce (Joyce et al. 2021), emphasizing that every technical decision is inherently political and is thus linked to values, power dynamics and potential inequalities. For example, a binary gender classification between males and females—that is often adopted in AI systems—obscures non-binary genders (Costanza-Chock 2020), illustrating that categories like gender and ethnicity are socially constructed and influenced by power relations (Carey and Wu 2023). Despite reluctance to take a political stance in AI development, not taking a stance is itself a choice (Benbouzid 2023). Based on these considerations, Kalluri (2020) advocates for AI fairness efforts directed towards altering power relations and involving vulnerable groups in co-creating AI systems, ensuring their voices and concerns are addressed, as discussed in Sect. 2.3.

Furthermore, while advocating for a more nuanced and context-specific interpretation of AI fairness, a sociological approach focuses on ideas such as a sufficiently in-depth, clear and informative explanation of AI systems' decisions (e.g., Borch and Hee Min 2022). This, however, can be in opposition to the more dominant fairness perspective of algorithmic fairness, which emphasizes numerical optimization. A sociological approach might emphasize, for instance, trading numerical metric scores for a more context-specific explanation of the system's decisions. Such tradeoff can be seen in having to pick whether to implement an interpretable, white-box AI model with lower benchmark scores, or an efficient black-box model that can only be explained to a lay person approximately and ad-hoc.

There is also emphasis on describing the AI system to different stakeholders. This should be considered in the context of, for example, the training of people who will be using the tool, but especially those most affected by what the outcome is—they not only need to know the decision and when the final decision is made in the first place, but also how to appeal, who is responsible, and so on (Starke et al. 2022).

2.3 Lay people's opinions on fairness

A growing body of literature on the topic of AI fairness emphasizes that the design and development of AI systems should follow a bottom-up approach starting from key stakeholders' ideas on fairness. This has been called the "participatory turn" in AI design (Delgado et al. 2023) and contrasts with the typically top-down way development teams work, which ends up building fairness from general concepts and well-established mathematical definitions. The top-down approach has been criticized for being expert-driven, with little consideration of the social context and the opinions and concerns of the actors who will use or be affected by the system (Greene et al. 2019).

In this regard, a pragmatic approach has been seen as a way of embedding lay people's views on fairness by building "ethics from people's lived experience, perceptions, narratives and interpretations" (John-Mathews et al. 2022: p. 946). Following this approach means dealing with the existence of pluralistic socio-technical (Bakiner 2023; Jasanoff and Kim 2009) and algorithmic imaginaries (Bucher 2016) that can influence how AI and its associated benefits and risks are perceived by different stakeholders. The first concept refers to institutionalized visions of a technological future as represented by corporations, social movements and professional societies, and the second to how lay people imagine, perceive and experience algorithms in their everyday lives.

Taking a participatory approach to the definition of fairness might imply for AI creators to deal with multiplicitous—and often conflicting—ideas on what constitutes a fair solution to a problem. For example, Lee et al. (2017) show the existence of multiple conceptions of fairness in the creation of an algorithm for the allocation of food donations; stakeholders have different opinions about what criterion would allow a fair allocation of resources (efficiency, equality, equity) and about what constitutes a fair procedure to assess the receivers' level of need. There are thus challenges on deciding whose values and whose ideas of fairness should be prioritized and implemented in an AI system, highlighting how every technical decision is a political decision that potentially favors some and penalizes others.

Moreover, participation in AI creation is influenced by power relations, which determine who gets to be part of the conversation, in what forms and to what extent. Depending on the level of power granted to the stakeholders in the decision-making process and their level of involvement, Delgado et al. (2023) distinguish four forms of participation: consultation, when stakeholders are involved to improve the user experience; inclusion, when stakeholders' values are incorporated into the design of the AI system; collaboration, when stakeholders decide on the system's features; and ownership, when stakeholders are involved throughout the project's lifecycle and decide on the actual scope of the system's creation. The preservation of power asymmetries between AI creators and the people affected by these systems can then negatively affect the real participation of stakeholders (Maas 2023), potentially leading to forms of "participation washing" (Sloane et al. 2022).

Many surveys investigate non-experts' hopes and concerns on the use of AI. Abuse of personal data, AI mistakes, unclear accountability, potential discriminations, and the fear of being controlled and manipulated are frequently cited concerns (BEUC 2020). On the other hand, improved services and management of everyday lives, efficiency, and cost savings are often listed among the benefits of AI systems (Pallett et al. 2024). Attitudes towards AI can be influenced by each person's level of algorithmic awareness (Gran et al. 2021), which can also lead to new forms of digital divide between those who possess the skills to critically interact with the systems and those who do not. Araujo et al. (2020) show that online self-efficacy—the perceived ability to protect personal data—is associated with higher expectations of the usefulness and fairness of automated decision-making and lower perceived risks. Furthermore, forms of algorithmic resistance (Bonini and Treré 2024) and self-help strategies to reduce the risks associated with the use of AI systems (Kappeler et al. 2023) show that non-experts are far from being passive recipients of technology and are finding ways to exercise their agency and play an active role in shaping and interacting with AI systems.

The inclusion of non-experts' opinions, values and concerns adds complexity to the creation of AI systems and requires AI creators to also address aspects that are difficult to operationalize, as we will show in Sect. 4.2, where we present the results of the involvement of vulnerable groups in our research.

3 Methodology

The research presented in the following subsections explored the ideas and opinions on AI fairness of experts from both technical and social sciences, as well as lay people, to see if and how they differ. The analysis of these perspectives on the same topic was aimed at finding points of contact and key differences, hereby forming a basis on how to reconcile different opinions and views. Experts from different disciplines were involved to identify clashing approaches that could complicate interdisciplinary work. On the other side, lay people contributed to the discussion by expressing their opinions and concerns on the use of AI systems in everyday life, bringing their own perspective to the table.

The study received ethics approval in February 2023 from the Ethics Committees of the University of Bologna and of the Centre for Research and Technology Hellas (CERTH). The research activities were conducted as part of the MAMMOth project, and were centered around three use cases. The first aims to bring fairness considerations when profiling people based on financial data (e.g., for credit scoring) and focuses on financial exclusion concerns, which are typically tied to social exclusion. The second aims to debias authorization decisions that leverage computer vision data to make predictions. For example, one sub-objective is to retain low false acceptance rates when authenticating people by comparing a picture with their identification document while providing similar rates for failing to identify members of different sensitive groups. The third investigates biases that occur when analyzing relational networks, how they arise, and which mitigation strategies can be applied. A focal point are disparities between genders and ethnicities in academic citation and collaboration networks, especially in the science and engineering fields.

Information sheets were distributed to the participants of the interviews and the co-creation workshops prior to the activity and signed consent forms were collected. Appropriate information sheets were distributed together with the survey and respondents were asked to provide their consent before filling the survey.

Our study has a few notable limitations. Because we focus mostly on qualitative rather than quantitative aspects, we do not claim to be representative of the general trends of the whole field of AI fairness, but rather to explore some of the underlying structural logic that many discussions surrounding AI fairness are built upon. For instance, although we conducted a survey as part of our methodology, we do not claim for it to be representative, but rather to serve as a method to better capture some of the prevailing thought patterns regarding AI fairness discussions, as well as to generate some ideas for our co-creation workshop activities. Furthermore, our preconceptions as researchers should be kept in mind while familiarizing with our study, as our focus on alleviating the power imbalances and inequalities currently present in the AI field guided our choice of methods and analysis.

3.1 Interviews with experts

Between April and July 2023, we conducted 29 semi-structured interviews with experts studying the topic of AI fairness. 17 experts came from the social sciences—for the most part from sociology—, 10 from computer science, 1 had a background from both disciplines and 1 came from the legal domain. The aim of the interviews was to reconstruct the technical and sociological view on the topic of AI fairness and thus to identify points of contact and differences. Experts from the two domains were therefore asked the same questions, which focused on identifying their ideas on how to build fair AI systems.

Recruitment was carried out by asking the MAMMOth project partners to indicate possible experts on the topic from the academia and/or non-profit organizations dealing with the topic. Further contacts were collected with the snowball sampling technique. This process led to a diversified group of interviewees in terms of gender, provenance, and career level. The interviews were fully transcribed and a thematic analysis was performed with the support of the software NVivo, in order to identify the key themes that the interviewees suggested as important for achieving AI fairness. Following this coding process, we compared the themes mentioned by social and computer scientists in order to identify areas of conflict and agreement.

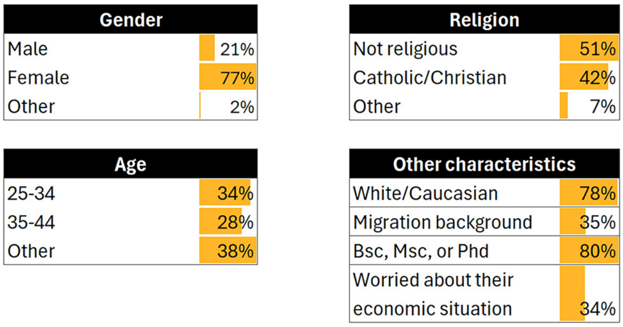
3.2 Gathering the opinions of lay people

As part of our investigation about lay people' perceptions on what fairness means to them, we co-organized a single survey and six co-creation workshops. The activities involved vulnerable groups at risk of discrimination by AI systems. Participants were recruited from partner organizations: Associaciò Fòrum Dona Activa 2010 working with women at risk of social exclusion including ethnic minorities and domestic violence victims; IASIS NGO working with ethnic minorities, young people and unemployed receiving social support; Diversity Development Group working with migrants, non-EU citizens and ethnic minorities; and Rijksuniversiteit Groningen that involved early career researchers belonging to the LGBTQI + community and/or ethnic and religious minorities.

3.2.1 Survey

The survey was circulated among vulnerable groups from partner organizations from the beginning of April 2023 to the end of May 2023. The survey was first designed in English and later translated into four different languages used by the groups that the survey was trying to reach: Spanish, Catalan, Greek, and Russian. The questionnaire was divided in five parts: (1) an investigation of each respondent's general level of awareness on AI, (2) their opinions on the use of AI in finance, (3) for ID verification, (4) in the academic citations and collaborations domain, and (5) the collection of the main demographic characteristics of the respondents, which are summarized in Fig. 1.

During the survey period, we collected 171 complete answers and 58 incomplete answers. The survey acted as an exploratory study into lay people's opinions on AI fairness.



Gender	
Male	21%
Female	77%
Other	2%

Religion	
Not religious	51%
Catholic/Christian	42%
Other	7%

Age	
25-34	34%
35-44	28%
Other	38%

Other characteristics	
White/Caucasian	78%
Migration background	35%
Bsc, Msc, or Phd	80%
Worried about their economic situation	34%

Fig. 1 Summary of survey respondent characteristics. Bars visualize representational imbalances by comparing each group's representation to the maximum (full bar) of the same group of characteristics

It did not aim to be representative, but to collect useful insights that could inform the co-creation activities. Only descriptive statistical analysis was thus performed. However, future probes should consider improving representation, especially if the end-goal is extracting quantitative fairness criteria.

At the beginning of the survey the respondents were asked to share their degree of familiarity with artificial intelligence. After this introductory question, they were provided with a definition of Artificial Intelligence and a textual description of the examples presented in Fig. 2—each referring to a different use case. The respondents were then asked to evaluate three AI use case scenarios: (a) a divorced man of close-to-retirement age being denied a loan due to his age being identified as a risk factor, (b) a black woman not being correctly identified by an AI identification system while comparing a selfie to an ID card picture, as AI systems often fail to identify black women, and (c) a researcher from a large university being prioritized in online search results over a researcher from a small university with similar research portfolio due to her not having as many international connections which affects the citation number.

3.2.2 Co-creation workshops

In May 2023 a total of six co-creation workshops were conducted on the MAMMOth project use cases (see Table 1). The topics that were explored during the workshops included, but were not limited to, the following ones:

1. The participants' opinions on the use of AI in the specific use cases.
2. What they thought a fair outcome would be when using these systems in real-life.
3. How they felt about the use of AI in specific use cases.
4. Their concerns about the use of AI in the use cases and the benefits that they identify.

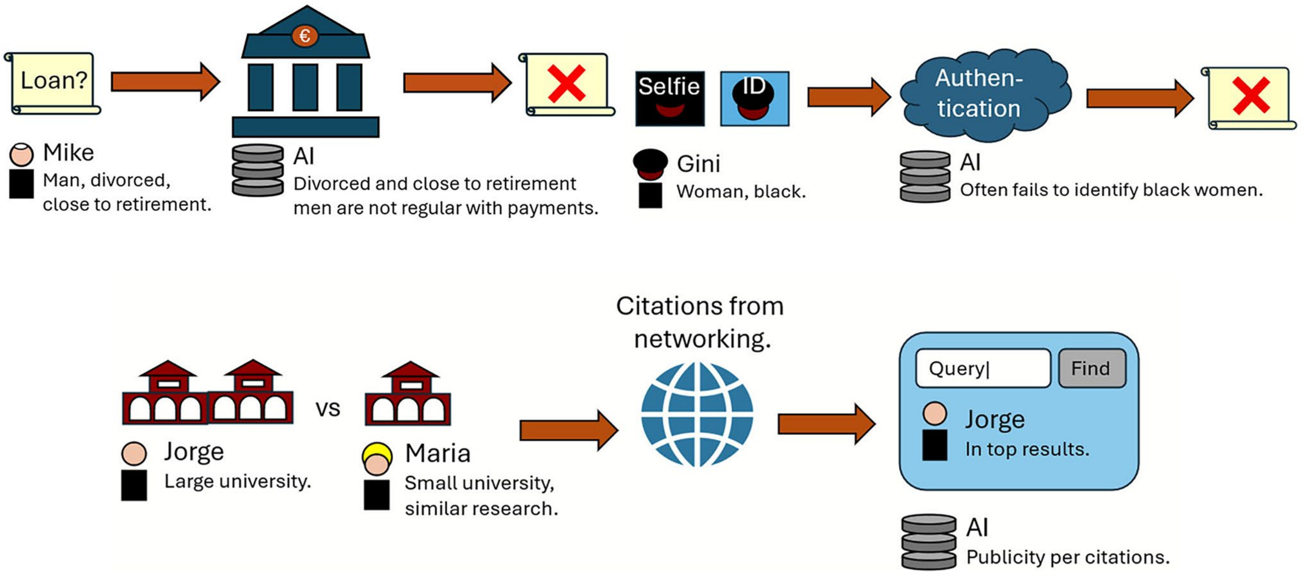


Fig. 2 The three AI use case scenarios evaluated by respondents

Table 1 Six workshops were conducted online and on-site in various locations

#	Use case(s)	Location	No. of participants
1	Academic citations and collaborations	Online	7
2	Finance, identity verification	Spain and online	17 on site and 2 online
3	Finance	Greece	20
4	Finance	Greece	4
5	Finance	Greece	4
6	Finance, identity verification	Online	5

5. Their previous experiences, if any, with the use of AI in specific use cases.
6. What groups of people they thought might be negatively affected by the use of AI in the use cases.

A brief training session on AI was conducted to provide the participants with at least a basic understanding of the topic and to familiarize the participants with basic terms and ideas. The training showcased real-life examples of how AI can reproduce discrimination and showed how AI works at a basic level. Also in this case, a thematic analysis was performed in order to identify the underlying ideas, meanings, and perceptions of the participants regarding AI fairness.

4 Results

We now present the main results of the above-described research activities. In Sect. 4.1, we analyze the interviews conducted with AI experts, and investigate possible points of contact between the approaches of social and computer sciences (Sect. 4.1.1). We thus try to build a common ground that can serve as a basis for discussion and interdisciplinary work on the topic. However, we also identify aspects that we believe to be harder to reconcile (Sect 4.1.2). For each excerpt, the number in square brackets indicates the identifier of the interview from which it was taken.

In Sect. 4.2, we present the opinions and concerns of the vulnerable groups involved in the survey and the co-creation workshops. These show how the involvement of non-experts in the development of AI systems creates new challenges in translating lay and general ideas into something that an AI system can understand. Nonetheless, it also broadens the perspective by including considerations that may have received little attention in algorithmic fairness research.

4.1 Interviews with experts

4.1.1 Points of contact

Fairness is context specific. Both computer and social science experts emphasized that there cannot be general definitions of fairness, because fairness is context specific.

Therefore, they stressed that any consideration regarding fairness needs to be contextualized as it can change in time and space. It was also mentioned that fairness is culturally dependent; what is considered fair may change when AI systems are used in new contexts. Furthermore, as one respondent noted: "when we consider issues of fairness, it's not about algorithmic fairness; it's about […] the fair creation of an algorithm within the context and the importance of these contextual factors" [I4], so understanding how an algorithm operates in a given particular context. There was then an agreement on the fact that AI solutions should be context/use case specific—referring to concrete models, concrete datasets, concrete applications, linked to theory/a specific harm/a specific community to avoid generalizations.

The interviewees also acknowledged the need for an interdisciplinary methodology that can be followed when approaching the problem of fairness in AI systems. Some interviewees also mentioned the importance of looking at the entire system's life cycle (model development, model application, etc.), in order to understand where and when biases and discriminations arise. This includes monitoring and maintenance after the system is deployed—e.g., periodically auditing the system and letting stakeholders evaluate it recurrently–, as new sources of unfairness could emerge after initial deployment.

Fairness is not only a technical problem. Many interviewees recognized that fairness is a multi-layered dimension that includes ethical, legal, social and technical aspects, among others. However, it was highlighted that fairness is often sold as a technical problem, which runs the risk of overlooking those aspects of the problem that escape technical solutions. Some of these non-technical aspects emerge clearly from the opinions of non-experts and are presented in the following subsection. A common opinion among the social scientists and the non-experts involved in our research is that AI should support and not replace human decision making, as human judgment should remain key. AI was defined by one of the interviewees as "a tool to assist humans in the decision, but where the human is given considerable agency" [I6]. Whilst this was recognized as a key fairness requirement, it is something that cannot be pursued via technical solutions, as it regards the fair implementation and use of an AI system in everyday life. However, it can still be supported by technical solutions like strategies to prevent users from being too willing to accept the results of the AI system. As mentioned by one of the interviewees: "it is better to have a less efficient machine, which keeps human attention high, rather than one that makes fewer mistakes and which we rely on completely" [I19].

Fairness regards the involvement of stakeholders. Both computer and social scientists mentioned that the creation of AI systems should be collective, deliberative, participatory—created through a constant dialogue with different stakeholders, including those who are not usually part of the debate due to linguistic, cultural, or status reasons.

Computer scientists stressed that the consequence is that developers might need to deal with different conflicting metrics. Fairness was defined by a computer scientist as a normative claim about the distribution of utility to different classes of individuals [I17]. The interviewee stressed that in AI systems there is the risk to treat one normative claim as the only one that counts as fairness by prioritizing some metrics over others—and thus some opinions over others. For this reason, social scientists highlighted that every decision concerning fairness in AI systems is a political decision that reflects certain values and potentially (re)produces power dynamics. In a situation where there are conflicting opinions on fairness, this is reflected in choosing to prioritize one perspective over another. An example of conflicting ideas was presented by one interviewee who has worked on the use of AI in finance and conducted some workshops with both lenders and people who could be potentially affected by the use of the system. The interviewee mentioned that, whilst lenders wanted AI to be systematic, the bank clients wanted AI to be empathetic and compassionate, considering their personal situation and history when making a decision.

The limits of fairness. Both computer scientists and social scientists stressed the importance of recognizing the limits of fairness; fairness was described by a social scientist as a "horizon" [I1], something that cannot be fully achieved, but that AI systems should tend towards. The interviewees stressed that, from a technical perspective, the limits of each system and of each metric should be made clear. Transparency and explainability were mentioned by both disciplines as key contributors to identifying these limits. Computer scientists focused on the use of Explainable AI (XAI) techniques to explain unfairness. Being clear on the limits of each system includes also for the system to signal cases of extrapolation, so when the system is making decisions based on data that go beyond its training.

On the other hand, social scientists highlighted the importance of creating processes that are understandable by different stakeholders, with different levels of knowledge of AI and how it works. The suggestion was to translate the processes to languages that are appropriate for different actors and while accounting for their level of information needs. In particular, it was mentioned that people affected by the use of AI should be told the reasons behind decisions, in order to guarantee their right to explanation. For example, in AI systems used in the finance sector, clients should be able to understand the reasons why they receive or are declined a loan. Transparency also allows us to understand what variables the system is using to decide and which have more weight in influencing the final decision. It was mentioned that it is also important to introduce the possibility of appeal if people impacted by the system don't agree with its output—and again appeal is possible only if the system is transparent.

4.1.2 Open tensions

In the analysis of the results, we have identified three methodological aspects that we believe to be more difficult to reconcile between the sociological and the technical approaches to AI fairness. First is the very process of finding *definitions of fairness* that can be translated into mathematical formulas. Whilst algorithmic fairness focuses on concrete definitions that can be expressed mathematically, in the sociological literature there are no definitions as such, but instead a focus on the contextual nature of the opinions on the topic and the conflicts that can arise from different perspectives. In fact, in social sciences, fairness is not something that can be fully computed and operationalized. For example, the experts mentioned that fairness regards not only the ability to build unbiased AI systems, but also the aim those systems are built for: unbiased AI systems can still be unfair when applied to unfair contexts or built for unfair purposes. Facial recognition was mentioned as an example that is inherently unfair due to its purpose and regardless of technical safeguards. Although these differences require more coordination work between disciplines, we believe that ways can be found to collaborate on these issues. For a proposal on an interdisciplinary approach to defining fairness, see the AI Creator's AI Fairness Definition Guide that was written as part of the MAMMOth project (Krasanakis et al. 2024).

A second aspect concerns the *foundations* of the aforementioned disciplines in approaching the study of AI fairness. Social sciences tend to have a *complexity-oriented* approach, aimed at problematizing the use of technology and its implications. This hinders engagement with a business world that is increasingly oriented towards techno-solutionism. Moreover, the social sciences approach clashes with how computer science tends to *simplify the world* so as to translate it to mathematical terms. For example, it was mentioned that, to account for a sociological perspective, computer scientists need to come to terms with the uncomfortable feeling that there are not always easy solutions to problems.

A third aspect regards the view on the *connection between the social and the technical aspects* of AI fairness. The interviews echo the literature consensus that fairness in computer sciences is seen primarily as a problem that can be approached through technical solutions. Such solutions include good quality training data and sampling strategies that are as balanced as possible with regard to sensitive characteristics; the need to train the system multiple times; the need for technical mitigation strategies. The experts reported that the main narrative in computer science describes unfairness as something present in the real world, so not something that an AI expert should worry about, as not part of their area of expertise [I12; I15]. However, from a sociological perspective, technical aspects are deeply intertwined with the social ones. As mentioned by one of the interviewees: “Fairness and unfairness cannot be identified without the social fabric in which the system is implemented” [I1], which includes an analysis of the power relations at play. This confirms why an interdisciplinary approach is needed in order to shed light on this interconnection and include “the decades of discussion about fairness in social sciences” [I11].

4.2 Lay people’s perceptions

In this subsection the results from our survey and workshops are provided. Most survey interviewees (87%) declared at least some knowledge of AI. The workshops, however, are closer to the reality of lay people. Although prior to the introductory presentation only a few participants were familiar with the current state and influence of AI in society, after the presentation the participants realized that they encounter narrow AI on a daily basis and use its services regularly, such as Google Maps, translation services, Spotify, etc. Some participants were surprised that AI tools are already utilized in high-stake systems such as banking. All these interactions once again showcase that, although non-experts may not be familiar with the technical side of AI, a large amount of consequences of AI use fall on their shoulders, whether with or without their knowledge.

4.2.1 Benefits of using AI

Most participants agreed that AI can bring some benefits in daily life and can make things more efficient. For instance, in workshops on finance, benefits like improved customer experience, freeing staff from repetitive tasks and fraud detection were named, among others.

Some people believed that AI systems can never truly be unbiased because of non-objective data, while others stated that objectivity could vary depending on the training data. The *negative to moderate* view was more prominent. Showcasing the importance of a fair training data set, one group stressed the importance of making sure that the training datasets ensure enough diversity in terms of age, gender, ethnicity, and religious symbols. Generally, while the potential benefits of AI in identity verification, such as cost and time efficiency, were recognized, these were overshadowed by significant concerns about privacy, security, discrimination, and fairness.

4.2.2 Fairness concerns

Although most participants agreed that AI has some benefits, in the end a more nuanced perception of AI systems emerged. In addition to the benefits, it was mentioned that AI systems should be thoroughly tested to make sure they don't (re)produce biases and forms of discrimination, and that AI needs regulations, shared standards and classifications.

In the workshops the participants recognized that AI systems will *"never be perfectly fair"*. However, the need to neutralize the downsides of systemic racism, discrimination and marginalization need to be emphasized. For instance, one workshop participant noted the discomfort of being evaluated by AI, stating, "the main feeling is that it's not fair that AI is judging. Because I am a migrant, even if I am equal to another person in all parameters, I will get a lower rating just because of my migration status, it is not fair" [E1]. Despite the acknowledgement from lay people that AI systems can, in fact, bring significant benefits to the everyday experience of both businesses and customers/consumers, the complex reality of balancing the benefits and the drawbacks while focusing particularly on the often complex personal context of the decision remains one of the key takeaways from both surveys and workshops.

Out of the three cases provided in the survey, the identity verification case was perceived as the most unfair. This could be explained by the fact that this case is more connected to a personal characteristic (i.e., skin color), whereas the other two examples (loan request and academic citation) focus more on characteristics which are usually considered less innate and on which opinions could be more mixed. For instance, identity verification feedback focused on the need for diverse training data, on the fact that the final decision should be made by a human and that the system should not be used if it's possible that it will make mistakes. As one participant noted, "Statistics is not about fairness, but about the majority and/or probability" [E4]. On the other hand, the feedback from the other two cases focused on how the university/research system works as a whole, that AI should support the process and a human being should be making the final decision, that the situation should have some flexibility and so on.

In other words, the other two workshop cases were more related to discourses on meritocracy, personal achievement, effort, etc. as opposed to protected characteristics such as gender or ethnicity. Discrimination on the later characteristics can sometimes be more visible, identifiable or more direct. Notably and interestingly, however, one respondent felt comfortable with AI decision-making precisely due to their demographic status: "I feel OK, because I am an educated white young male" [E3]. Furthermore, it's worth mentioning that the positive aspect that was recognized the most in all of the three cases was the potentially saved time due to increased efficiency and productivity.

Particularly in the workshops about identity verification, participants were largely uncomfortable with AI's involvement due to fears about data leakage. Statements such as, "don't like it, no matter AI or human", "strange feeling", "can't relax" and "afraid" illustrate such apprehension. This illustrates *the problem of privacy*, which is not only based on an abstract concern with the person's private affairs, but also is based on a possibility of an actual data breach occurring.

To summarize, results from our survey and workshops indicate a few important points about the lay people's perceptions of a fair AI. Firstly, the final decision should be *transparent, flexible and continuously monitored*, and also such decisions should not be left to the automated system itself but rather should be in the hands of a human being. Furthermore, concerns cover the (lack of) diversity of training and testing data and how AI might be reproducing already existing inequalities and even further marginalizing already marginalized groups, especially in regards to ethnicity, gender and other innate or protected characteristics. Additionally, specifically when it comes to banking practices and finance, there was a cynicism expressed that the profit-oriented businesses may not prioritize fairness. In relation to ID verification, it was added that people from lower socio-economic background could be additionally disadvantaged, for example, due to lower phone or camera quality working as a proxy for socioeconomic status. In the end, although some benefits of using AI were mentioned, the more negative attitudes prevailed.

It is worth mentioning that during one of the workshops a participant confessed to using a chatbot (ChatGPT) for their responses in the initial part of the workshop. This incident showcases not only the ubiquity and pervasiveness of narrow AI tools in everyday life, but also the complexity of researching the role of AI systems and tools in the life of its users.

5 Discussion: negotiating fairness

The results described in Sect. 4 give a hint into the complex nature of the concept of fairness when applied to AI systems. The results show that fairness encompasses many different—and sometimes conflicting—aspects. The analysis shows the subjective nature of fairness, meaning that different stakeholders may have different perceptions of what is fair and what is not, and that there is no objective definition or implementation of fairness. For this reason, which idea of fairness is implemented in an AI system should be something negotiated among different actors and based on compromise. However, as suggested also by the interviewees, processes of negotiation are often affected by power and information imbalances. Characteristically, some actors are often left out of the conversation, vulnerable people do not have enough space to express their opinions, and negotiation tends to favor the ideas of actors with more power (Delgado et al. 2023).

For this reason, we believe that the above-described results reveal an overarching theme: that of power relations. As shown in the results section (Sect. 4), this topic has been mentioned on multiple occasions by the experts as the underlying condition that affects the involvement of multiple voices in the design of fair AI systems. As argued by our interviewees, in order to build fair AI systems, their construction process itself must be fair and inclusive. This entails an effective involvement of all stakeholders, even those most difficult to include, and a real negotiation between them, where everyone can find space to express their opinions. Power is in fact something that is enacted, negotiated, and contested in everyday interactions, rather than a fixed resource held by specific entities. Power relations are produced and maintained through processes of participation, knowledge, and resistance, highlighting the ongoing negotiation of power (Gaventa 1980). Power is then both enabling and constraining, in an ongoing process (Giddens 1979): current power relations affect the decision on who gets to be part of the conversation around AI fairness—what disciplines and what stakeholders—but can be either confirmed or challenged in social practices and interactions.

We therefore argue that working towards fairer AI systems is only possible by altering current power relations, both between disciplines—questioning the current hegemony of computer science and putting other disciplines on an equal footing—and between experts and lay people—considering the latter equally entitled to express their opinions. In this section, we compare the results gathered through our research activities with the current literature on algorithmic fairness. We try to identify gaps in the literature that should be filled in order to move towards a lay-socio-technical view of AI fairness (Sect. 5.1). We believe that moving in this direction requires two processes of negotiation: among disciplines, and between experts and lay people. We therefore discuss prospective strategies that can help these processes of negotiation (Sects. 5.2 and 5.3, respectively). With the subsections below, we therefore hope to contribute to the strand of literature focusing on how to build AI systems that are able to question current power relations (e.g., Kalluri 2020; Corbett et al. 2023; Birhane et al. 2022; Delgado et al. 2023) by providing concrete directions in this sense.

5.1 The table of fairness concerns

We start by summarizing the concerns voiced by computer and social science experts, as well as lay people. Some of these concerns are already recognized by algorithmic fairness research described in Sect. 2.1. The points of contact among disciplines are presented in Table 2 as broad categories to think about when negotiating fairness. They also reveal the importance of fairness negotiation, as different actors represent different concerns tied to their roles in relation to AI systems (e.g., creators, users, stakeholders), and other actors should start actively considering those concerns. We further organize points of contact in accordance with each concern's place in a software project's lifecycle (Ruparelia 2010—depending on the exact development methodology, these steps may be revisited down the line).

A basic, yet core idea connecting different flavors of AI fairness(es) is discrimination mitigation: every user should be treated without bias, discrimination and preconception. What differs between the approaches, however, is *how* this goal should be reached, and what kind of fairness paths are prioritized by different stakeholder groups or in the literature. Take explainability as an example. For computer scientists, it often means a post-hoc approximation of the black-box model that checks whether some numerical metric is sufficiently covered or not. For the sociologists or lay people, however, it refers to more "useful" knowledge: how and based on what information the decision was made, the reliability and accuracy of such explanation, what the user can do about it, etc.

When it comes to algorithmic fairness, discrimination mitigation often takes the form of research and fairness software tools that tailor definitions of fairness and respective measures of bias to specific settings being examined. Popular fairness libraries like AIF360 (Bellamy et al. 2019) and FairLearn (Bird et al. 2020) provide a wide breadth of algorithmic solutions to a specific subset of well-studied problems. Emphasis is placed on catering to different predictive tasks, such as optimizing classification, regression, or ranking. This approach focuses on algorithmic understanding of fairness metrics and similar numerical leads, but does not consider the broader sociopolitical context of these problems that's required for a well-rounded and fair system. The overlooking of, for instance, the Human-in-the-loop (HITL) concern (i.e., when a human expert is supervising/intervening with AI's activities) is an instance where the lack of social background is found in both the algorithmic fairness literature and amongst computer scientists (though the computer science experts we interviewed are somewhat more aware of the social context than the algorithmic fairness literature is).

One exemplary attempt to remedy the numerical fixation has been the algorithmic recourse. Algorithmic recourse solutions, such as counterfactual explanations, attempt to counteract the broader sociopolitical negligences, but are lacking due to their focus on merely providing information without taking into account societal power imbalances—e.g.,

Table 2 Comparison of fairness concerns between researchers, experts, and non-experts

Concern	Literature	Gathered Feedback		
	Algorithmic fairness	Computer scientists	Sociologists	Lay people
Design				
Analysis of inequalities and power relations			✓	
Creation of metrics	✓	✓		
Data quality (e.g., fairness diversity)	(✓)	✓		✓
Involvement of stakeholders	(✓)	✓	✓	
Interdisciplinary approach	(✓)		✓	
Legal compliance or standards	(✓)	✓		✓
Negotiation between stakeholders (e.g., fairness vs accuracy))		✓	✓	
Transfer of definitions between domains	✓			
Unquantifiable concerns		✓	✓	✓
Development				
Algorithmic implementation of fairness	✓			
Dependence on context/use case		✓	✓	✓
Discrimination mitigation	✓	✓	✓	✓
Robustness	✓	✓		
Maintenance				
Algorithmic recourse	(✓)		✓	
Explainability	✓	✓	✓	✓
Flexibility				✓
Human-in-the-loop and appeals	(✓)		✓	✓
Ongoing monitoring	(✓)	✓	✓	

(✓) refer to emerging but still not global adoption of the respective concerns

the fact that the actions needed to change the AI decision's outcomes can be simply not possible to carry out, like "increasing one's income to get a loan".

Such oversights can be supplemented by a sociological account of fairness. Sociologists criticize various technical fixations, such as those on data quality or that of legal compliance, as superficial and non-performative, while also emphasizing the deconstruction of inequalities and power relations. By adding a sociological account of a situation, the purely technical account becomes somewhat more contextualized in the social reality, encouraging an interdisciplinary view.

Although such a combined sociotechnical and expert view of the fairness question tackles many issues, it is subject to the threat of trying to find a one-size-fits-all solution. To counteract this, the engagement and active involvement of *all* of the stakeholders should be emphasized. Table 2 shows that the lay people's need for flexibility was not mentioned by the experts. This indicates that any type of expert knowledge can be too generalized, non-context specific and lacking the nuance of the reality on the ground. Even when the technical and the sociological approaches are fused together, this does not in any way guarantee that the actual needs to achieve fairness for the users are taken into account.

And yet, concerns such as robustness (the ability of the system to withstand unexpected errors or conditions), while crucial to computer science experts, are often overlooked by lay people themselves. This suggests that the people directly influenced by AI's decisions might remain ignorant of injustices done to them due to unforeseen real-world circumstances or malicious actors gaming the system. This, again, brings one back to the expert approaches and shows that an active negotiation of what a fair system should look like must take place between the various stakeholder groups. Only then can the chances of missing critical AI fairness conversation details be minimized. As the concept of fairness gains more steam, the emphasis on negotiation helps to resist the commodification of the concept into just another metric to optimize.

Although this discussion is in no way exhaustive, creating fair AI systems means making compromises or sacrifices on the side of the privileged between technical efficiency and context-specific concerns (particularly of system users). We reiterate that, during system creation, the concerns of lay people can be too easily ignored due to the power imbalances between those making technical decisions and those affected by them. Thus, we encourage the involvement of *all* of the stakeholders, especially the marginalized ones.

5.2 Takeaways: vocabulary-conversations-education

This section examines possible ways to establish and maintain a process of negotiation between computer scientists and sociologists. We identify three main aspects that can favor interdisciplinary work on the topic of fairness in AI.

Shared vocabulary and boundary objects. Different disciplines often use a different terminology or the same terms with different meanings. For example, what is commonly referred to as bias in the computer science domain is analyzed in terms of inequalities and power relations in sociology. Bias is broadly considered, in computer science, as an inherent imbalance in data that skews predictions towards certain outcomes without that being the original intent. In terms of AI fairness, problematic biases are those that skew predictions against certain individuals based on undesirable characteristics. The aim of fairness-aware algorithms is then to remove such biases from AI systems.

By contrast, sociology studies how biases originate from pre-existing social inequalities and systems of oppression at the expense of marginalized groups (Carey and Wu 2023), in particular for those biases identified as "societal biases" or "historical biases" by computer scientists (Zajko 2022). This different approach on biases has consequences on how fairness is examined in the sociological field, transforming it to a pursuit of "structural social change". Fairness, then, includes the effort not to replicate the power imbalances and inequalities that already exist in society (Weinberg 2022). For this reason, some authors (Kong 2022; Giovanola and Tiribelli 2022) distinguish between weak or negative fairness, which pursue the technical debias of AI systems, and strong or positive fairness, which is about actively challenging oppression. If biases are deviations from the "ground truth" (Benbouzid 2023), but the ground truth—society—is deeply unequal, debiasing AI systems can correspond to reproducing the inequalities found in the society (Zajko 2022). If biases are considered more broadly as something undesirable, this opens further possibilities in promoting social change.

In this context, building a shared vocabulary was suggested by various interviewees as a first step in order to favor interdisciplinarity. This can represent a "boundary object" (Star and Griesemer 1989) that can support the cooperation among disciplines. Boundary objects are concrete or abstract objects that simultaneously inhabit different social worlds, although sometimes with different meanings. These mismatches become "problems for negotiation" (*ivi*, p. 412) among disciplines. However, boundary objects have at the same time enough in common to allow communication between worlds.

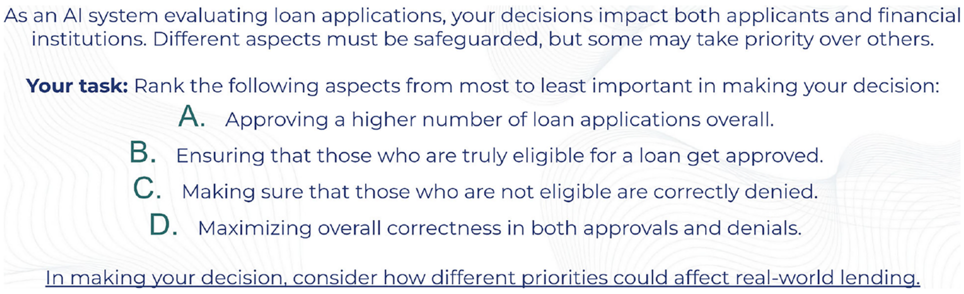
Having uncomfortable conversations. Such conversations should include recognition of missing expertise with regards to certain aspects of the domain. This means challenging a broad tendency to blindly accept certain epistemic viewpoints, which create definitions of fairness that are *known* to be inadequate (e.g., like the 4/5ths rule that can be an indication of bias but whose absence is not an indication of fairness), yet are still being employed on a large scale.

Following previous research on the topic (Majchrzak et al. 2012; Kou and Harvey 2022), we can distinguish two ways to approach knowledge gaps. The first practice is about "traversing" knowledge boundaries and involves the search for a common ground through the discussion of the deep level assumptions on which each discipline is based and their inherent differences. This approach pursues the creation of "bridges between knowledge domains" (Kou and Harvey 2022: p. 1356). The second practice is about "transcending" knowledge boundaries and regards the creation of new knowledge through the combination of the expertise of different disciplines. However, it doesn't involve the discussion of each discipline's assumptions.

A number of interviewees mentioned the importance of having "uncomfortable conversations" [I7] among experts from different disciplines, resembling those traversing practices mentioned above. However, the two approaches are not mutually exclusive and are often used interchangeably in interdisciplinary work. Moreover, traversing processes do not necessarily lead to consensus and are again the result of a negotiation process involving potential conflicts. These unresolved conflicts can represent useful moments for reflection (Hirsbrunner et al. 2024). Even the shared knowledge and representations that are the results of these processes "contain at every stage the traces of multiple viewpoints, translations and incomplete battle" (Star and Griesemer 1989: p. 413).

Education. Another point that was stressed multiple times by interviewees as a way to support interdisciplinary work was that of education. Interviewees highlighted that university courses rarely foster interdisciplinarity and disciplines are organized as separated silos. This aspect makes interdisciplinary work less easy to pursue. The necessity to have a two-way education was then stressed by interviewees with the aim of building a mutual construction of sociology and technology in the context of AI systems. It means supporting the entanglement of sociology in technical complexities and of technology in social complexities. If we consider again the example of biases mentioned above, sociology can highlight the origins of biases and the consequences of technical choices on people's lives, whilst computer science can show what is feasible to achieve with technical solutions. Communication then requires a set of shared knowledge that serves as a basis for interdisciplinary work. The interviewees talked

Fig. 3 Options for deciding which predictive quantities should be compared



As an AI system evaluating loan applications, your decisions impact both applicants and financial institutions. Different aspects must be safeguarded, but some may take priority over others.

Your task: Rank the following aspects from most to least important in making your decision:

A. Approving a higher number of loan applications overall.

B. Ensuring that those who are truly eligible for a loan get approved.

C. Making sure that those who are not eligible are correctly denied.

D. Maximizing overall correctness in both approvals and denials.

In making your decision, consider how different priorities could affect real-world lending.

about this topic mentioning the necessity to "smooth out the corners of disciplines" and "meet halfway".

5.3 Translating lay people's opinions to algorithmic fairness

In this section we analyze how to strengthen and encourage dialogue between lay people and the two expert groups, namely computer scientists and sociologists, with the end goal of algorithmically quantifying and monitoring certain fairness concerns. To this end, existing technical works either generate definitions of fairness based on desired statistical properties or express statements about fairness in formal logics, i.e., using mathematical expressions like "A and B" and "A implies B". However, these approaches typically ignore the opinions of lay people.

Our call for encouraging the incorporation of lay people's opinions in the creation of AI can be complex. One way would be through software engineering pipelines of extracting, refining, and following fairness specifications. This is already costly and time-consuming for specifications that benefit business owners (not to mention the refinement across several system versions if Agile[4] development is followed) and is made doubly complex by involving lay people with neither technical nor domain expertise.

We followed a viable methodology within the MAMMOth project that builds on our insights. This first lets computer scientists select from applicable measures of algorithmic fairness and split those into simpler building blocks. Then, a negotiation with social scientists turns blocks into non-mathematical options for which stakeholders' opinions can be gathered. A common building block already mentioned is which base measures of predictive quality should be equalized between demographic groups. We exemplify options for the financial use case in Fig. 3; this asks for ranked preference between the mathematical quantities of positive rate (A—pr), true positive rate (B—tpr), true negative rate (C—tnr) and accuracy (D—acc).

To fully flesh out the example, gathered opinions for the financial and identity verification use cases are presented in Fig. 4. Following the methodology suggested by Demšar (2006) we employed a Friedman test to verify that at least one measure is perceived differently with statistical significance with p value < 0.01 and a Nemenyi post-hoc test to check which differences are significant at p value < 0.05 (this uses a greater value because post-hoc tests are known to be less powerful). This led us to select—approximate—tpr equality as the primary concern to safeguard for the financial use case as one that is preferred against others with statistical significance. For identity verification, we consider all three of tpr, tnr, acc as of primary importance. Despite equal representation in AI outputs (equal pr) being the most studied concern in the algorithmic fairness literature, it was less important than alternatives for stakeholders in both examined contexts, highlighting the dissonance between theory and opinions on the ground.

Another option is to consider many definitions of algorithmic fairness that capture a broad range of concerns (e.g., through the FairBench library that computes many measures and presents them systematically—Krasanakis and Papadopoulos 2024). Sufficient parameterization also enables fine tuning to the social context. For example, in fairness definitions where causal models are generated by domain experts by creating conceptual links between high-level terms, modeling parameters are learned from societal observations. In simpler scenarios, ad-hoc mathematical definitions may introduce hyperparameters to choose anew in each setting, such as the maximal accepted bias before systems are considered unfair. In the above example, a parameterization of how to compare base measures (e.g., by computing differences between groups) and bias thresholds can be obtained from stakeholder feedback (Krasanakis and Papadopoulos 2025).

Regardless of the strategy followed, involving lay people in technical decisions requires the transcription of technical concepts to intuitive interpretations. This is hindered by the fact that explainable AI algorithms tend to be outperformed by black box counterparts or are less accurate approximations of the latter (though this is disputed by Rudin 2019). An alternative is to devise interpretable fairness evaluation.

[4] The agile manifesto: https://agilemanifesto.org/principles.html (Accessed 17/10/2024).

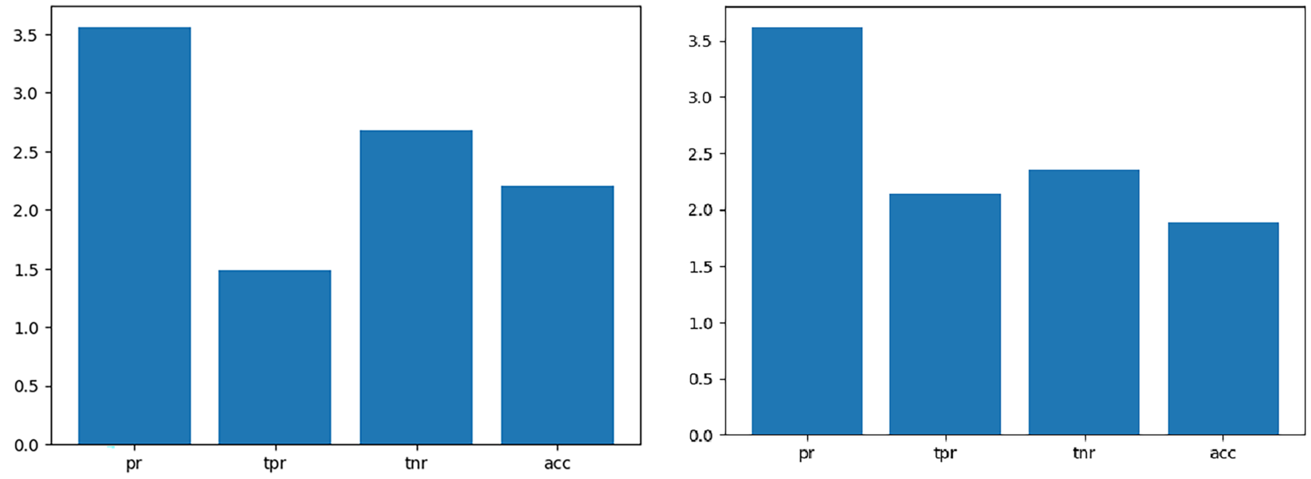


Fig. 4 Average rank (lower is better) of which quantity to equate between groups obtained by 48 stakeholders for the financial (left) and identity verification (right) use cases. Lower ranks indicate more preferred measures, and rank differences over 0.67 are considered statistically important with *p* value 0.05 per the Nemenyi test

For example, parameters of high-level fairness statements may be determined by gathering and aggregating the opinions of lay people through social science processes. Machine learning can also be employed to generalize gathered opinions to similar new settings. The aspects of fairness that end up being assessed through mathematical measures need to be constantly examined and revisited when (new) unfair scenarios are discovered. Continuous monitoring may come at steep engineering costs, and therefore well-structured processes should be introduced to simplify the renegotiation of fairness. The AI Creator's AI Fairness Definition Guide (Krasanakis et al. 2024) discusses what such an iterative process may look like.

A final challenge lies with how to adopt qualitative characteristics that cannot be measured from the real world. For example, accounting for the individual circumstances of loan recipients could be desired, but such information is both multifaceted and not standardized. Processes that convert qualitative characteristics into quantitative ones could involve checklists and step-by-step descriptions, and one needs to look at the humans applying such processes. In general, involving humans in the interpretation of AI system predictions is often necessary to satisfy fairness sensibilities, where those humans must be trained to understand the systems to the degree that is possible, and must be considered trustworthy themselves, for example by being held partially accountable. Creating and maintaining AI that accommodates human interpretation is costly, for example due to requiring comprehensive user interfaces. It also prevents systems from being deployed at scale, in which case humans in the loop can only look at aggregate information and may still miss individual circumstances.

Finally, it is crucial to acknowledge that education plays a key role not only in the dialogue between computer scientists and sociologists—as stated in Sect. 5.2—but also in the promotion of a real inclusion of lay people in the processes described above. Lay people often possess an intuitive sense of justice, yet may overlook the underlying structures and societal factors influencing justice. Education plays a key role in illuminating these frameworks, fostering a more informed and inclusive dialogue about justice in AI. Sociological approaches are especially valuable in this regard, as they shed light on power dynamics and societal structures that shape our perceptions of justice.

6 Conclusions: rebalancing power relations

The multiple layers that define fairness call for more diversity in the way the issue is approached and tackled. As we have shown, this applies both to the disciplines involved, calling for more integration of social sciences in the way these systems are developed, and to the stakeholders that are part of the negotiation process. This process of negotiation implies accepting and learning how to deal with potential conflicts—both between disciplines and between stakeholders. It also requires promoting awareness on fairness and discrimination to make sure that non-experts can be an active part in the process.

As mentioned by the interviewees, there cannot be a conversation about fairness without addressing power imbalances when fairness is related to ensuring access, representation and equity to different disciplines and stakeholders. Power relations are at the core of sociological analysis and they represent a crucial lens to see the world in the discipline. The results show that the process of negotiating fairness requires a redefinition of the power relations at play to ensure a variety of perspectives and include those voices that are often not part of the discussion. We have shown how this process of negotiation entails power relations both between disciplines, where computer science tends to have more power than social sciences in determining the direction of technological development, and between experts and lay people, where people from vulnerable communities are often left out of the conversation.

In this article, we have tried to identify some strategies in order to pursue this goal. However, power relations are inherently dynamic and subject to continual change, rather than being static or permanently balanced. Consequently, strategies must be flexible and responsive. They require ongoing renegotiation and adaptation of power relations to respond to evolving contexts and new insights. This perspective underscores the necessity of continuous reassessment and adjustment of power dynamics to promote equity and inclusivity in the ongoing process of AI development.

Acknowledgements The authors would like to express gratitude to all the members of the MAMMOth project "Multi-Attribute, Multimodal Bias Mitigation in AI Systems" for their input on the preparation of the interview guide, survey and material for the co-creation workshops. We are particularly grateful to: Anna Pujol Ferran, Eva Serret and Virginia Carreras from Associació Fòrum Dona Activa 2010, Elli Nikolakopoulou from IASIS NGO, Jogaila Vaitekaitis from Diversity Development Group, Dagmar Heeg and Dali Fekete from Rijksuniversiteit Groningen for their contribution to the organization and conduction of the co-creation workshops and/or the preparation and circulation of the survey; Evren Yalaz from Trilateral Research Limited and Stavroula Rizou from the Centre for Research and Technology Hellas for their support in the preparation of the Ethical Committees approval requests; Dagmar Heeg and Dali Fekete from Rijksuniversiteit Groningen for the preparation of the training material for the co-creation workshops.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Author contributions This article is based on the joint efforts of all authors. However, M.G. wrote: Sect. 2.3; Sect. 4.1 and its paragraphs; the introduction to Sect. 5; Sect. 5.2; the Conclusions. E.K. wrote: Sect. 2.1; Sect. 5.3. K.G. wrote: the introduction to Sect. 2; Sect. 2.2; Sect. 4.2 with its paragraphs. The introduction was written jointly by M.G. and E.K. Paragraph 5.1 was written jointly by E.K. and K.G. All other parts were written jointly by the authors. The interviews were conducted by M.G. The survey and co-creation workshops were designed by M.G. with the support of the people included in the acknowledgements. Data analysis was performed by M.G. All authors read and approved the final manuscript.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. This research work was fundedby the European Union under the Horizon Europe MAMMOth project, Grant Agreement ID: 101070285. UK participant in Horizon Europe Project MAMMOth is supported by UKRI grant number 10041914 (Trilateral Research LTD).

Data availability A more detailed overview of the results is publicly available at: https://mammoth-ai.eu/wp-content/uploads/2023/11/d1.1-v1.0.pdf. Raw data of this paper, which includes data from workshops, interviews, and surveys, are not publicly available to preserve individuals' privacy under the European General Data Protection Regulation. Following the research ethics approval obtained from the ethics committees of CERTH and the University of Bologna, personal information collected through these research activities is pseudonymized by removing participants' names and all other identifying features. Only pseudonymized data is used in research products, including this paper. To minimize the risk of re-identification and comply with the ethics approval, access to the pseudonymized data is restricted to the research team and only available upon reasonable request, under conditions that guarantee the confidentiality and anonymity of the participants.

Declarations

Conflict of interest The authors declare no competing interests.

References

Krasanakis E, Gibin M, Rizou S (2024) The AI Creator's AI Fairness Definition Guide. 1st Edition. https://github.com/mammoth-eu/FairnessDefinitionGuide. Accessed 16 October 2024

Krasanakis E, Papadopoulos S (2024) Towards Standardizing AI Bias Exploration. In: Proceedings of AI bias: Measurements, Mitigation, Explanation Strategies, CEUR, 3744

Angwin J, Larson J, Mattu S, 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 16 Oct 2024

Araujo T, Helberger N, Kruikemeier S, De Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc 35(3):611–623. https://doi.org/10.1007/s00146-019-00931-w

Bakiner O (2023) Pluralistic sociotechnical imaginaries in artificial intelligence (AI) law: the case of the European Union's AI act. Law Innov Technol 15(2):558–582. https://doi.org/10.1080/17579961.2023.2245675

Bankins S, Formosa P, Griep Y, Richards D (2022) AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. Inf Syst Front 24(3):857–875. https://doi.org/10.1007/s10796-021-10223-8

Barabas C, Doyle C, Rubinovitz JB, Dinakar K (2020) Studying up: reorienting the study of algorithmic fairness around issues of power. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (FAT* '20). Association for Computing Machinery, New York, pp 167–176. https://doi.org/10.1145/3351095.3372859

Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: limitations and opportunities. MIT Press, Cambridge, MA

Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Zhang Y (2019) AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4/5):4–1. https://doi.org/10.1147/JRD.2019.2942287

Benbouzid B (2023) Fairness in machine learning from the perspective of sociology of statistics: How machine learning is becoming scientific by turning its back on metrological realism. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency (FAccT '23). Association for Computing Machinery, New York, pp 35–43. https://doi.org/10.1145/3593013.3593974

BEUC (2020) Artificial Intelligence: what consumers say. Findings and policy recommendations of a multi-country survey on AI. https://www.beuc.eu/sites/default/files/publications/beuc-x-2020-078_artificial_intelligence_what_consumers_say_report.pdf. Accessed 16 Oct 2024

Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K (2020) Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft, Tech Rep MSR-TR-2020-32. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf. Accessed 16 Oct 2024

Birhane A, Isaac W, Prabhakaran V, Diaz M, Elish MC, Gabriel I, Mohamed S (2022, October) 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). Association for Computing Machinery, New York, pp. 1–8. https://doi.org/10.1145/3551624.3555290

Bonini T, Treré E (2024) Algorithms of resistance: the everyday fight against platform power. MIT Press, Cambridge

Borch C, Hee Min B (2022) Toward a sociology of machine learning explainability: human–machine interaction in deep neural network-based automated trading. Big Data Soc 9(2):20539517221111361

Bucher T (2016) The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. Inf Commun Soc 20(1):30–44. https://doi.org/10.1080/1369118X.2016.1154086

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency (FAccT '18). Association for Computing Machinery, New York, pp 77–91

Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data Min Knowl Discov 21:277–292. https://doi.org/10.1007/s10618-010-0190-x

Carey AN, Wu X (2023) The statistical fairness field guide: perspectives from social and formal sciences. AI Ethics 3(1):1–23. https://doi.org/10.1007/s43681-022-00183-3

Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2021) The zoo of fairness metrics in machine learning. arXiv abs/2106.00467. https://doi.org/10.21203/rs.3.rs-1162350/v1

Corbett E, Denton R, Erete S (2023) Power and public participation in AI. In Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization (EAAMO '23). Association for Computing Machinery, New York, Article 37, pp 1–13. https://doi.org/10.1145/3617694.3623228

Costanza-Chock S (2020) Design justice: community-led practices to build the worlds we need. MIT Press, Cambridge

Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven

Delgado F, Yang S, Madaio M, Yang Q (2023) The participatory turn in AI design: theoretical foundations and the current state of practice. In: Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization (EAAMO '23). Association for Computing Machinery, New York, Article 37, pp 1–23. https://doi.org/10.1145/3617694.3623261

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Dodge J, Liao QV, Zhang Y, Bellamy RK, Dugan C (2019) Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th international conference on intelligent user interfaces (IUI '19). Association for Computing Machinery, New York, pp 275–285. https://doi.org/10.1145/3301275.3302310

Dong Y, Ma J, Wang S, Chen C, Li J (2023) Fairness in graph mining: a survey. IEEE Trans Knowl Data Eng 35(10):10583–10602. https://doi.org/10.48550/arXiv.2204.09888

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012, January) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). Association for Computing Machinery, New York, pp 214–226. https://doi.org/10.1145/2090236.2090255

Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St Martin's Press, New York

Fenu G, Galici R, Marras M (2022) Experts' view on challenges and needs for fairness in artificial intelligence for education. In: Rodrigo MM, Matsuda N, Cristea AI, Dimitrova V (eds) Artificial intelligence in education. AIED 2022, vol 13355. Lecture notes in computer science. Springer, Cham, pp 243–255. https://doi.org/10.1007/978-3-031-11644-5_20

Gaventa J (1980) Power and powerlessness: quiescence and rebellion in an Appalachian valley. University of Illinois Press, Champaign

Giddens A (1979) Central problems in social theory: action, structure, and contradiction in social analysis. University of California Press, Berkeley

Giovanola B, Tiribelli S (2022) Weapons of moral construction? On the value of fairness in algorithmic decision-making. Ethics Inf Technol 24(1):3. https://doi.org/10.1007/s10676-022-09622-5

Gran AB, Booth P, Bucher T (2021) To be or not to be algorithm aware: a question of a new digital divide? Inf Commun Soc 24(12):1779–1796. https://doi.org/10.1080/1369118X.2020.1736124

Greene D, Hoffmann AL, Stark L (2019) Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Hawaii international conference on system sciences, pp 2122–2131

Hirsbrunner SD, Tebbe M, Müller-Birn C (2024) From critical technical practice to reflexive data science. Convergence 30(1):190–215. https://doi.org/10.1177/13548565221132243

Jasanoff S, Kim SH (2009) Containing the atom: sociotechnical imaginaries and nuclear power in the United States and South Korea. Minerva 47:119–146. https://doi.org/10.1007/s11024-009-9124-4

John-Mathews JM, Cardon D, Balagué C (2022) From reality to world. A critical perspective on AI fairness. J Bus Ethics 178(4):945–959. https://doi.org/10.1007/s10551-022-05055-8

Joyce K, Smith-Doerr L, Alegria S, Bell S, Cruz T, Hoffman SG, Nobel SU, Shestakofsky B (2021) Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. Socius 7:2378023121999581. https://doi.org/10.1177/2378023121999581

Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature 583(7815):169

Kappeler K, Festic N, Latzer M, Rüedy T (2023) Coping with algorithmic risks. J Digit Soc Res 5(1):23–47. https://doi.org/10.33621/jdsr.v5i1.130

Karimi AH, Schölkopf B, Valera I (2021) Algorithmic recourse: from counterfactual explanations to interventions. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAccT '21). Association for Computing Machinery, New York, pp 353–362. https://doi.org/10.1145/3442188.3445899

Khan FA, Manis E, Stoyanovich J (2021) Translation tutorial: fairness and friends. In: Proceedings of the ACM conference on fairness, accountability, and transparency (FAccT '21). Association for Computing Machinery, New York, pp 1097–1105

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). Association for Computing Machinery, New York, pp 485–494. https://doi.org/10.1145/3531146.3533114

Kou CY, Harvey S (2022) A dialogic perspective on managing knowledge differences: problem-solving while building a nuclear power plant safety system. Organ Stud 43(9):1355–1378. https://doi.org/10.1177/01708406211061864

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. Advances in neural information processing systems. In: Proceedings of the 31st international conference on neural information processing systems (NIPS '17). Curran Associates Inc., Red Hook, pp 4069–4079

Lee MK, Kim JT, Lizarondo L (2017) A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In: Proceedings of the 2017 CHI conference on human factors in computing systems (CHI '17). Association for Computing Machinery, New York, pp 3365–3376. https://doi.org/10.1145/3025453.3025884

Li B, Qi P, Liu B, Di S, Liu J, Pei J, Zhou B (2023) Trustworthy AI: from principles to practices. ACM Comput Surv 55(9):1–46. https://doi.org/10.1145/3555803

Luchs I, Apprich C, Broersma M (2023) Learning machine learning: On the political economy of big tech's online AI courses. Big Data Soc 10(1):205395172311538. https://doi.org/10.1177/20539517231153806

Maas J (2023) Machine learning and power relations. AI Soc 38(4):1493–1500. https://doi.org/10.1007/s00146-022-01400-7

Majchrzak A, More PH, Faraj S (2012) Transcending knowledge differences in cross-functional teams. Organ Sci 23(4):951–970

Marcus G (2020) The next decade in AI: four steps towards robust artificial intelligence. arXiv:200206177

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Annu Rev Stat Appl 8(1):141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Morozov E (2013) To save everything, click here: the folly of technological solutionism. Public Affairs, New York

Nakao Y, Stumpf S, Ahmed S, Naseer A, Strappelli L (2022) Toward involving end-users in interactive human-in-the-loop AI fairness. ACM Trans Interact Intell Syst (TiiS) 12(3):1–30. https://doi.org/10.1145/3514258

Noble SU (2018) Algorithms of oppression: how search engines reinforce racism. New York University Press, New York

Ntoutsi E, Fafalios P, Gadiraju U et al (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdiscip Rev Data Min Knowl Discov 10(3):e1356. https://doi.org/10.1002/widm.1356

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Penguin Books, London

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Pallett H, Price C, Chilvers J, Burall S (2024) Just public algorithms: mapping public engagement with the use of algorithms in UK public services. Big Data Soc 11(1):20539517241235868. https://doi.org/10.1177/20539517241235867

Parraga O, More MD, Oliveira CM et al (2023) Fairness in deep learning: a survey on vision and language research. ACM Comput Surv. https://doi.org/10.1145/3637549

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Ruparelia NB (2010) Software development lifecycle models. ACM SIGSOFT Softw Eng Notes 35(3):8–13. https://doi.org/10.1145/1764810.1764814

Sartori L, Theodorou A (2022) A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics Inf Technol 24(1):4. https://doi.org/10.1007/s10676-022-09624-3

Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y (2019) How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society (AIES '19). Association for Computing Machinery, New York, pp 99–106. https://doi.org/10.1145/3306618.3314248

Sloane M, Moss E, Awomolo O, Forlano L (2022) Participation is not a design fix for machine learning. In: Proceedings of the 2nd ACM conference on equity and access in algorithms, mechanisms, and optimization (EAAMO '22). Association for Computing Machinery, New York, Article 1, pp 1–6. https://doi.org/10.1145/3551624.3555285

Star SL, Griesemer JR (1989) Institutional ecology, "translations" and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. Soc Stud Sci 19(3):387–420. https://doi.org/10.1177/030631289019003001

Starke C, Baleis J, Keller B, Marcinkowski F (2022) Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. Big Data Soc 9(2):20539517221115188. https://doi.org/10.1177/20539517221115189

Weinberg L (2022) Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches. J Artif Intell Res 74:75–109. https://doi.org/10.48850/arXiv.2205.04460

Zajko M (2022) Artificial intelligence, algorithms, and social inequality: sociological contributions to contemporary debates. Sociol Compass 16(3):e12962. https://doi.org/10.1111/soc4.12962

Zhang J, Bao K, Zhang Y, Wang W, Feng F, He X (2023) Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. In: Proceedings of the 17th ACM conference on recommender systems (RecSys '23). Association for Computing Machinery, New York, pp 993–999. https://doi.org/10.1145/3604915.3608860

Zhou J, Chen F, Holzinger A (2022) Towards explainability for AI fairness. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W (eds) xxAI—beyond explainable AI. xxAI 2020, vol 13200. Lecture notes in computer science. Springer, Cham, pp 375–386. https://doi.org/10.1007/978-3-031-04083-2_18

Krasanakis E, Papadopoulos S (2025) Interpretable and Adjustable Definitions of AI Fairness using Fuzzy Logic (Preprint). Zenodo. https://doi.org/10.5281/zenodo.16995442