

Equitable AI Audits: evaluating the evaluators in today's world

Monique Munarini
University of Pisa
Pisa, Italy
monique.munarini@phd.unipi.it

James Brusseau
PACE University
New York, USA
jamesbrusseau@gmail.com

Lorenzo Angeli
University of Trento
Trento, Italy
lorenzo.angeli@unitn.it

Abstract

Artificial Intelligence (AI) technologies have triggered substantial transformations across various sectors, yielding benefits and rasing concerns. A significant issue pertains to the unequal distribution of AI advantages among different groups, raising questions about diversity and fundamental rights. This paper proposes equity as a translational mechanism intersecting ethical, legal, and technical dimensions in AI governance. By tracing the evolution of AI regulation from corporate ethical guidelines to forthcoming legal frameworks, this paper sheds light on the growing emphasis on auditing mechanisms to mitigate bias and discrimination, distinguishing equity from the more often-used term fairness. While fairness emphasizes fair processes, equity prioritizes equitable outcomes. We argue that equity offers a value that surpasses the limitations of fairness and adopts a holistic viewpoint, providing a comprehensive framework for tackling socio-technical challenges in AI development and deployment, and preserving fundamental rights. In proposing a definition of equity that draws inspiration from feminist perspectives rather than traditional philosophy, we highlight equity's contextual nature across diverse global regions and cultures. This framing lets us propose equity as a non-universal value that can be pivotal in curating datasets, fostering community empowerment, and engendering a sense of belonging within the diverse communities affected by the AI tools used in decision-making processes. We further argue that ethics-based auditing serves as a crucial link between technical requirements and legal norms, offering insights into the complex socio-technical dynamics of AI development. This paper critically examines the integration of equity in AI audits, assessing whether current audits adequately address non-discrimination and fundamental rights or merely scratch the surface. With this paper, we wish to contribute to scholarly discussions on AI's role in preserving fundamental rights, and stress the importance of interdisciplinary collaboration in navigating AI's multifaceted challenges.

CCS Concepts

• **Applied computing** → *Law, social and behavioral sciences*; • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics** → Computing / technology policy.



This work is licensed under a Creative Commons Attribution International 4.0 License.

ICEGOV 2024, October 01–04, 2024, Pretoria, South Africa
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1780-2/24/10
https://doi.org/10.1145/3680127.3680140

Keywords

equity, reliable AI, AI governance, accountability

ACM Reference Format:
Monique Munarini, James Brusseau, and Lorenzo Angeli. 2024. Equitable AI Audits: evaluating the evaluators in today's world. In *17th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2024), October 01–04, 2024, Pretoria, South Africa.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3680127.3680140

1 Introduction

The intensive use of AI systems to support the decision-making process and to make it more agile has brought numerous benefits and risks. As AI increasingly integrates into our daily lives, the need to ensure reliable and ethical AI systems that respect fundamental human rights has become paramount [13]. This focus on AI's capabilities and its use as a service [12] also advanced the discussion about the challenges posed by biases, misinformation, and misuse of such technologies [7]. Previous research highlighted the urgency to regulate AI systems to prevent harm and safeguard individual rights, recognizing that laws are the only enforceable measure to effectively prevent AI misuse [41]. Vinuesa et al. [49], in particular, pointed out that artificial intelligence can be a powerful ally in the overall achievement of the UN Sustainable Development Goals but, conversely, it can also be a tool to increase differences among nations, genders and industries. As legislation regulating AI across the world are still under development, ethical principles are being sought as the basis for creating methodologies that can systematically monitor these AI-based tools.

The literature recognises AI systems as socio-technical entities, encompassing both technical elements (such as algorithms, code, and data) and social components (including developers, deployers, and corporations) [2, 18]. In other words, AI is by nature multidisciplinary and, to allow efficient collaboration between experts from different fields, it is important to have a common understanding of key values. In this paper, we propose equity as an ethical value because of its ability to permeate different sectors in the AI ecosystem [42], working as a bridge between the legal and technical domains. In 2018, Floridi [19] argued that "hard ethics" are the guidelines for the creation of legal frameworks because hard ethics relates to moral dualities between what is right and wrong when developing legal norms. Nonetheless, what Floridi called "soft ethics" provides the basis for the development of compliance mechanisms, as it is related to feasibility between what is prescribed in the law and how to enforce the norm, as a form of post-compliance ethics. Stahl [42] further brought the notion of ecosystem to the field of AI considering that, similarly to a biological ecosystem, during the AI lifecycle, this system of interdisciplinary interactions, many diverse stakeholders are interconnected in complex relationships. Auditing and monitoring processes play a critical

role in ensuring that AI technologies are developed and deployed responsibly. This acknowledgment has led to a growing reliance on third-party auditing processes, aiming to independently assess and verify AI systems' conformity with legal and ethical requirements. However, while the regulatory frameworks provide a foundation, the effectiveness of these auditing mechanisms in upholding ethical AI remains an open question. In conclusion, the confluence of AI's societal impact, regulatory efforts, existing violation of fundamental rights connected with AI deployment, and the reliance on auditing mechanisms to implement binding norms sets the stage for a comprehensive exploration of the role of equity in fostering reliable AI and safeguarding fundamental rights.

Methodologically, this research is not based on a systematic and exhaustive review of everything that has been published on the topic, following other scholars researching topics with plural meanings [44]. Rather, it is a critical equity-scoping of how equity is perceived among key stakeholders in AI Governance, aiming to build a working definition common to all fields from the different understandings of the broader category of equity. With this paper, we aim to shed light on how equity can be used as a translational mechanism that intersects with ethical, legal and technical fields of AI. This systemic intersection aims to understand how this definition benefits the development of monitoring and auditing methodologies for AI.

In Section 2, and in particular in 2.1, we introduce the reasons why we pursued research on the topic of equity and the relationship with reliable AI. In Section 2.2 we briefly describe how equity is being used in the legal, ethical and technical domain within the AI ecosystem. Followed by our proposition of working definition of equity in Section 2.3. In Section 3, we suggest how the presented equity definition can be used to assess current auditing mechanisms and what is the benefit in doing so to build reliable AI. Finally, in Section 4, we draw our concluding remarks on the importance of equity as a value for AI development.

2 Equity

Our research started by seeking to understand what differentiates equity from the much-more-common category of fairness. In this section, we will discuss these differences (that we synthesise in Table 1), discussing how equity is seen by ethical, legal and technical scholars within the AI ecosystem.

2.1 Why Equity?

Our research started from acknowledging a growing hype about using fairness across different domains in AI. There is countless academic research connecting fairness and AI Governance (e.g., [27]). Some papers use fairness and equity interchangeably [10, 48], though we argue the two need to be distinguished. We can trace the origin of the conception of fairness in the Western thought back to Aristotle [4], stating that equals should be treated equally and unequals unequally, according to their inequalities. On the non-technical side of the AI ecosystem, fairness is often used as a synonym for equality and non-discrimination, mostly connected with the legal principles under the same name. However, the technical side of the AI ecosystem assigns a completely different meaning to fairness, usually associating it with techniques for bias mitigation. Ntoutsi and colleagues show how, in over fifty years, there have been more than twenty distinct definitions of fairness [34]. As "ethics-by-design" developed by Brey and Dainow [9] associates fairness with mitigation of bias and illegal discriminatory treatment, fairness also became heavily discussed within developers and policymakers. These tensions around the term "fairness" make it difficult to identify solutions that in fact have a holistic perspective able to solve social-technical issues in the AI development and deployment, as bias, for example.

As for equity, instead, we can recognise two fundamental theoretical positions that govern its application, and that set it apart from fairness.

The first, as we show in Fig. 1, frames equity as being about outcomes, and fairness as being about processes. Equity begins by conceiving a desirable outcome (e.g., inclusiveness and equality across society) and then lets that vision dictate a process (e.g., the evaluation of resumes by an AI tool). Conversely, fairness starts from defining the most desirable process (i.e., treating equals equally, and unequals unequally within the decision domain), and letting the application of that process dictate the outcome (e.g., in hte same example of an AI tool evaluating resumes, deciding who is hired). Elaborating the example, proponents of equity will start with the outcome they desire, which may be achieving a diverse composition in an organisation, and then set a process to facilitate that result. Since statistical methods in computing research have been shown to be tractable for this purpose, by adjusting weights in the algorithms that order and label inputs, the resulting outputs can be shaped to accord with the original vision [38]. Advocates of fairness will instead concentrate on justifying the hiring process (e.g., giving all applicants similar ability tests in similar conditions), and then let the equality of the hiring process justify the outcome, regardless of the composition of those who are finally hired. Again, in equity, the outcome dictates the process; in fairness, the process dictates the outcome.

The second way to distinguish equity and fairness sees equity as being about collectives, and fairness as being about individuals. In equity, individuals exist first as members of the larger society: who an individual is, is determined by the larger society that holds them. Contrastingly, a fairness views posits that societies exist as collections of individuals: the guiding ideals and values defining a society are the cumulative ideals and values of its members. Here, equity starts from a vision of society, which in turn dictates how individuals in that society will be treated; fairness starts from a vision of how individuals in society will be treated, and that dictates the society that will emerge. Again, an example on hiring can be helpful. Proponents of equity will profile and justify a certain shape for their (collective) workforce. The profile may be drawn in terms of gender or race or some other factor, but that is a secondary consideration: what matters is that the collective profile determines how individuals will be understood and treated, it decides who will be hired and promoted. For equity, it follows that the way an individual exists is determined for that individual by the larger group. The individual may not even know if they are (individually) defined by their race, gender, capacity for the job, industriousness, etc. By contrast, a fairness view sees individuals in society as treated equally or unequally in proportion to their professional capacity for the task that the individual applied for.

Table 1: A summary comparison of equity and fairness

Value	Scope	Focal point	Approach	Definition
Equity	Outcome-focused	Collective	Aiming an equitable society dictates how individuals will be treated	Understanding and providing to the ones who need the necessary resources to belong in a particular community
Fairness	Process-oriented	Individual	Tailored process that will define how individuals will be treated and this will shape a fair society	*Ethics*: "Treat equals equally and unequals unequally" [4] *Law*: A process or decision is considered fair if it does not discriminate against people based on their membership to a protected group (EU perspective) [3] *Computer Science*: Typically treated in a reductionist way, seeing fairness only as a quantifiable statistical metric [10]

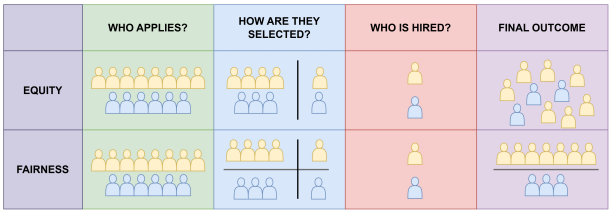


Figure 1: A comparison between Equity and Fairness in a hiring process. Here, we wish to highlight how Equity emphasises providing equitable *outcomes*, while Fairness emphasises providing a fair *process*.

What makes an individual who they are is nothing more or less than their individual ability to do that task. The resulting equal (and unequal) treatment in hiring and promoting across multiple individuals will subsequently dictate the composition of the larger workforce. Again: equity means the collective sets the rules for individuals; fairness means individuals setting rules for themselves will subsequently define the composition of the collective. More broadly, equity is the conviction that a good society defines what it means to treat individuals justly, while fairness is the conviction that treating individuals justly defines what it means to have a good society. This can impact in including marginalised voices instead of just escalating hegemonic perspectives [7].

Having proposed these two views we can argue that equity, then, is a term with similar understanding across disciplines. Choosing a term that can be perceived by different audiences with the same meaning is thus is essential if we want to build common understanding — and eventually, reliable AI.

2.2 Equity in the ethical, legal and technical AI fields

The journey to regulate artificial intelligence evolved from early corporate ethical guidelines to a complex web of international standards and now maturing into legal frameworks [15].

The initial phase saw the emergence of corporate ethical guidelines, as tech giants and industry leaders began to acknowledge the need for responsible AI development and deployment [35]. However, an analysis of these guidelines showed that they were not taken into consideration in the decision-making process in the design of AI systems, mostly because they were not enforced within the very corporations who developed them [23]. Globally, key international stakeholders have undertaken efforts to regulate AI technology, reflecting the need for comprehensive frameworks to navigate its complex societal implications. In the present context, AI regulation has matured into a phase of legislative action from nations worldwide [15] enacting laws specifically addressing AI. The European Union has taken a pioneering role with the AI Act, aiming to ensure AI systems' trustworthiness within its member states. Meanwhile, countries like Canada, China, and the United States are also introducing their own AI-specific regulations, each tailored to their particular societal and economic contexts [1]. Concurrently, the OECD, UNESCO, and even the G20 have issued non-binding ethical guidelines, reflecting a consensus on the importance of AI governance at the international level [20]. The possible overlaps between ethical principles and laws cannot raise the assumption that they are equivalent or that the concepts can be used interchangeably [41], but it is important to ensure that there is a common understanding when applying concepts present in both fields.

The AI Principles from the OECD consist of five main principles in which equity is related to social justice, and addresses that AI systems must be developed and harnessed in a manner that reveres human rights. This involves creating AI technologies that empower individuals by enhancing their decision-making capabilities, all while respecting core ethical considerations. In the UNESCO's AI principles, laid out in the "Recommendation on the Ethics of Artificial Intelligence", values are introduced as influential guiding principles that shape policies and laws. The specified values inspire positive actions and serve as the basis for principles. Furthermore, principles elaborate on the underlying values, making them more practical and applicable in policy formulation and implementation. In this sense, "ensuring diversity and inclusiveness" and "living in a peaceful, just, and interconnected society" are presented as values while "fairness and non-discrimination" and "multi-stakeholder and adaptive governance and collaboration" are the related principles. In both cases, the term equity is mentioned as a concept related to inclusiveness, diversity and intersectionality as part of a process

to achieve multi stakeholder engagement and to avoid any type of discriminatory action. The UN Interim Report Governing AI for Humanity released in December 2023 by the AI Advisory Body [8] highlights established grounds to develop a risk assessment framework that can uphold global AI Governance. Equity, together with ethics, human rights, gender, and societal impact was selected as one of the cross-cutting issues to be examined in the work of the Advisory Body and one of the core principles for the development of an AI Global Framework.

In the legal domain, equity is usually associated with the notion of social justice and what is fair and just in the administration of justice [43]. The EU Charter of Fundamental Rights [46] brings in its titles III and IV the rights under equality and non-discrimination. Although the term "equity" itself is not explicitly mentioned, the concept is closely related to rights of non-discrimination and equal treatment found in Article 21 (Non-discrimination) and Article 20 (Equality before the law). Furthermore, the Anti-Discrimination Law within the EU encompasses Directives 2004/113/EC and 2996/54/EC, aimed at combating discrimination and promoting equality. These Directives address various forms of discrimination, including gender and age-based discrimination, ensuring equal treatment across diverse contexts. The European Parliament Resolution 2017/3016 on empowering women and girls through the digital sector [47] is an important milestone in setting guidelines to increase diversity, mostly related to the gender gap in the ICT Sector through gender mainstreaming. In acknowledging these differences, gender mainstreaming sets a series of strategies related to policy-making to introduce a gender-sensitive approach in each phase of the policy development. It also addresses intersectionality and diversity in considering different gender perspectives, beyond the binary categorization of men and women [21]. These Directives underscore the EU's commitment to fostering a society free from discrimination and upholding fundamental rights as prescribed in the EU Charter.

Equity's ethical basis, instead, traces back to John Rawls' philosophy, as articulated in his Theory of Justice [39]. There, Rawls proposes a counter to the standard notion of fairness as defined by Aristotle. He argues that an objective person placed in the role of distributing a societal good would, in fact, diverge from the Aristotelian notion of equal apportioning. Instead, for Rawls, a deliberative approach tilts the distribution toward those who have the least. Funding for education, for example, may be tilted to favour schools with the poorest performance instead of being distributed equally. Or, taxes may be lowered for a society, but the cuts should be designed to benefit those with the lowest incomes. Because Rawls was interested first in political philosophy, the societal goods he envisioned aligned well with the kinds of benefits governments distribute, including educational opportunities and funding, infrastructure construction, taxes, and similar. However, the specific content of the distribution is less important than the logic governing the distribution, and this logic is the core of what has been referred to as "solidarity," and what today is often named by the word equity. The specifics of the logic's development — which include discussions of Rawls' "original position" and his "veil of ignorance" — stand outside the scope of this paper, but the results concentrate in what Rawls called his "difference principle" [39] which stipulates that inequalities may be allowed only to the extent that they most benefit the worst-off. In an unequal world, this translates into the general rule that the maximum advantage should go to the least advantaged, and that rule can be applied across domains, from governmental decisions about distribution to private company decisions about hiring.

There are already some studies in Computer Science directed to equity and questioning the limitations of fairness, in the sense that only focusing on fairness metrics might not ensure a fair deployment of the tool depending on the context [40]. The group Queer in AI [37] shared a case study about participatory methods in AI designing and deployment to address intersectional marginalisation of communities that are not under the radar when fairness initiatives are designed. In this specific case, their achieved outcomes were equity, inclusion, community and role models by using community-based participatory action research as a decolonizing epistemological framework to interrogate power and privilege. In recent years, the term equity was boosted in the policy field by the Biden-Harris administration that incorporated equity in their executive orders. This action raised a lot of questioning if equity would exacerbate discrimination, and discussions about the definition of the term itself [30]. Also, inclusion and diversity are broader societal challenges, usually associated with equity, that have been emerging as a central challenge to AI development [18]. Again, the notion of intersectionality, inclusiveness and diversity connected with the term equity is very clear from different sources of knowledge within the AI ecosystem [42].

2.3 Building a common definition of equity

As the main contribution of this research, we want to propose a definition of equity that can be operationalized across domains. To do so, we chose to use feminist thinking to support this task. Feminism is rooted in the acknowledgement of power imbalances in society and ways of redressing these differences to achieve equity [5]. Rather than relying on a technological fix, the idea to search in society what can be understood by equity intends to understand what is required to empower marginalised groups that are not being embraced by current solutions. This is in line with Haraway's [24] idea of bringing diverse voices in the discussion about technology. Haraway argues [25] that *translation* is a ground for power-sensitive conversations without loosing its rational critical, and always interpretive features. A terminology that can bridge different fields of knowledge is essential to build effective strategies to mitigate possible harms considering that AI systems are social-technical entities, as already highlighted in this paper. For this reason, a common definition of equity would have this translational function to facilitate cooperation among different stakeholders. Moreover, from the previous sections it is possible to note how equity is already present in key domains in one way or another and a grounded definition would just reinforce this liason as illustrated in Fig. 2. In the figure, AI Governance can be understood as any effort to avoid the escalation of harms from the deployment of AI systems. These efforts can arise from regulation, self-regulation and advocacy, as discussed in Section 2.2, and enable us to build a definition of equity that will incorporate elements from ethics, law and tech.

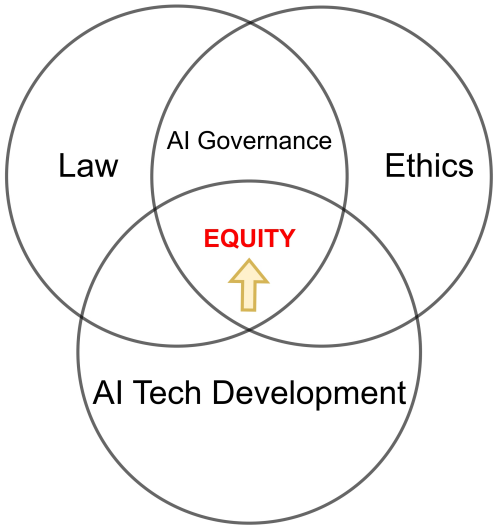


Figure 2: Illustration of the translational function of equity in the intersection of different domains in the AI ecosystem

Simone de Beauvoir's play, "Les bouches inutiles" or "Who shall die?" [6] explores the theme of equity and the importance of inclusiveness in decision-making within a community set in a besieged 14th-century village. The village Council, dominated by elite men, decides to sacrifice vulnerable groups for survival, sparking dissent. A female character challenges this, leading the villagers to unite and fight for collective freedom beyond age and gender distinctions. Beauvoir is not focusing on the exceptions and how individuals should be treated differently, but in how to include the widest range of diverse members in a community. Creating a group of marginalised members in the community such as women, children and elderly, reinforces the logic of oppression from one group to the other, raising the dilemma of who should live and who should die. In the AI decision-making process, the same outcomes in choosing just the hegemonic group as default for society is very present [36]. Treating people differently because they are part of a specific group might be *fair*, but it is not *equitable*, because this is the confirmation that they do not belong to a particular community. The marginalised group becomes the "others" that must, according to a given regulation, participate (see final outcome on Fig. 1). It is "acting like an insider and feeling like the outsider" [29].

Wajcman [50] argues that the idea of a socio-technical ecosystem reinforces the need to contextualise the meaning, effects, and values embedded in technologies, as they vary by culture and country. Moreover, she suggests looking at educational principles as a key resource for an equitable society. Within the AI ecosystem, the notion of equity is usually associated with educational practices and, most recently, with ethical and equity issues in collaborative learning [11]. The UNESCO report on generative AI in education [45] highlights equity as a key indicator when evaluating whether generative AI used for educational purposes exacerbates disparities in access to technology and educational resources, further deepening inequalities. In other words, to be able to participate and share an inclusive and diverse environment, those who are usually left behind need to feel that they belong to their community. Paulo Freire [22] discusses the importance of equitable education, exploring the relationship between power, oppression, and education. He proposes an educational approach that includes marginalised communities, acknowledging that they might have different beginnings in their educational path due to social inequalities, but the sought outcome is the same. To achieve this common goal, he advocates for an educational system that ensures equal access to quality education and resources regardless of social background. The holistic comprehension of societal needs is summarised in the idea that "nobody is superior to anyone else" [22]. Equity, in the sense of Beauvoir's and Freire's work, is related to the idea of bringing everyone on board without exception.

As Minow [30] described, equity is more than bringing people to the table: it is about allowing them to meaningfully participate in the decision-making process. To Freire, the inequalities that "popular kids" might face outside school would not be reproduced in the school environment: within the gates of the educational institution, they would receive the same quality education as "privileged kids". The focus should be on addressing the underlying structural inequities and working towards equalising educational outcomes for all students. Equity, therefore, is not universal, as communities are not. As community varies across different regions globally, with different cultures, equity is a non-universal value, which does not ignore or disconsider protected groups, but goes beyond axes of social difference to address intersectionality and achieve inclusiveness through diversity.

In the AI ecosystem, an equity approach can foster a thoughtful process of curating datasets that will be used in the development of AI systems. This can impact by including marginalised voices instead of just escalating hegemonic perspectives [7]. Freire [22] already flagged the exclusion of false participation that manipulates contributors, centralising the decision-making process. The idea of including the "other" is present in Beauvoir's work when she argues that "woman is the other" and Freire when he claims that we became a society incapable of seeing the other, to assure the respect of their human rights. Audre Lorde [29], in criticising feminism for overlooking intersectionality, argued that ignoring the differences among women regarding races was a serious threat to women's joint power. In the case that a white woman, not acknowledging her privileges, defines a woman in terms of her solely experience, then women of colour become "other", "the outsider whose tradition is too 'alien' to comprehend" [29]. Without diving into the philosophical debate about the other, equity allows diversity and inclusion because it can address the needs of others so they can participate at the same level within the community they are inserted in [33].

Equity, in essence, ***involves providing the necessary resources to those who need them to foster a sense of belonging in a specific community***. This concept encompasses four key elements: (i) identifying those in need, (ii) determining the essential resources, (iii) understanding the target community, and (iv) achieving a sense of belonging within that community. A real example of how this equity definition could be used to solve issues related to an AI system can be helpful. Ali et al. [2] found that the AI-based optimisation process in Ad delivery tools from Google and Meta were replicating gross gender stereotypes and showing different job offers according to the gender of the user.

An equity approach needs to consider these four aspects. Again, going back to our example on hiring:

(i) Identifying those in need: Begin by identifying specific demographic groups that historically face barriers in recruitment processes. For instance, women, migrants, or individuals from underrepresented socioeconomic backgrounds may encounter challenges in accessing job opportunities. Consider how intersectionality could impact differently applicants by inviting those groups to the assessment.

(ii) Determining essential resources: Assess the resources needed to address the identified disparities. This could involve designing an AI algorithm that undergoes rigorous training on a diverse dataset that actively counters and corrects biases, provide bias awareness training to hiring managers and recruiters to identify potential pitfalls from the system considering the received applications. Take steps to ensure that job recommendations are not influenced by traditional gender roles or stereotypes.

(iii) Understanding the target community: Make sure that marginalised groups receive the job advertisements, and that they have the chance to apply to them. Engage with both job seekers and employers, collaborating with diversity and inclusion experts, gender equality advocates, and industry professionals to gain a nuanced understanding of the challenges faced by diverse communities in the job market.

(iv) Achieving a sense of belonging: When companies use tools to assist in hiring, offering personalised and unbiased job recommendations, they actively discourage the replication of gender stereotypes by ensuring that all users receive job offers based solely on their skills, qualifications, and preferences. Systems should strive to improve the delivery of job offers by providing detailed insights and guidance to users, fostering a sense of belonging by emphasising individual strengths and potential rather than conforming to expected roles.

This example serves as an illustration of the complexities involved in fostering equity within AI systems from a real case. It emphasises the necessity of adopting a multidisciplinary approach in the design and implementation of AI technologies aimed at achieving equitable outcomes for society. The key elements identified—identification of those in need, determination of essential resources, understanding the target community, and achieving a sense of belonging — highlight the importance of comprehensive strategies that go beyond mere technical considerations. This example not only brings attention to the potential pitfalls in AI applications but also highlights the potential impact of an equitable AI system on reshaping societal norms and preserving fundamental rights.

3 How to think about equity in AI audits

One notable aspect of the evolution of regulation around AI is the increasing focus on addressing bias and discrimination in AI systems through auditing mechanisms. Laws like the New York Bias Law [16] are designed to hold AI developers accountable for the potential biases inherent in their technology. The EU AI Act, instead, have provisions on fundamental rights impact assessment and risk assessments as compulsory measures to systems classified as high risks [14].

There are several types of accountability tools designed to support compliance with legislation and ethical guidelines during the AI lifecycle such as (i) certificates, (ii) standards, (iii) impact assessments, (iv) regulatory sandboxes, and (v) auditing and monitoring mechanisms [28]. Certificates (i), for instance, serve as official attestations that an AI system complies with specified requirements or ethical frameworks, promoting transparency and trust among users and stakeholders. Standards (ii), on the other hand, establish a common framework that outlines technical and ethical norms for AI development and deployment, fostering consistency, interoperability, and reliability across diverse applications. Impact assessments (iii) evaluate the potential ramifications of AI systems on societal, economic, legal, and ethical fronts. These assessments offer insights into potential risks and negative outcomes, enabling key changes in development and deployment strategies. Regulatory sandboxes (iv) provide controlled spaces for developers to experiment with innovative AI solutions while adhering to limited regulatory restrictions, thereby encouraging innovative strides within manageable risk boundaries. Auditing and monitoring mechanisms (v), finally, ensure that AI systems continue to uphold legal and ethical requirements over time. This ongoing evaluation involves periodic assessments, data audits, and performance reviews to promptly identify and address any deviations or concerns. The distinction between monitoring and auditing is highlighted by Hasan et al. [26], who state that monitoring processes are usually internal processes conducted during the development of AI systems in a non-independent way to provide feedback about the compliance of legal or ethical frameworks, while auditing processes are independent evaluations that can be conducted by third or second parties.

Ethics-based auditing goes beyond legal compliance [31], assessing the outcomes of an entity, not necessarily AI, according to ethical principles or norms. Ethics risk assessments within a second- or third-party auditing process are focused on the potential harmful effects that outcomes from an AI system might cause [32]. The idea of using ethics as a bridge between the technical performance of an AI system and the need to be compliant with legal norms is essential in addressing this holistic perspective of the AI ecosystem as attentive to complexities of socio-technical context [31].

Our discussion of equity, contextualised in the broader discourse of AI accountability, calls for widespread application of our proposed definition. This definition challenges prevailing monitoring and auditing methodologies, which often take ethical values as given, and instead of preventing rights violations, can inadvertently exacerbate them. Our hope is to enrich the discourse on equity in a holistic manner, genuinely advancing the development of ethical AI and hoping to open the discussion of how we may rethink some audit methodologies. The four elements of equity proposed in Section 2.3 offer guidance for how audits may craft tools for analysing and monitoring specific uses of AI-based systems. Our definition is not a mere checklist for the scrutiny of data and algorithms; it emphasises a meaningful participation of marginalised groups throughout the design process, particularly given their heightened vulnerability. Incorporating our approach necessitates a paradigm shift from focusing solely on the AI tool's performance to contemplating the socio-technical environment that empowers individuals within their communities. Equity should not merely be mentioned to justify operations, but operationalized in practice.

This concept of equity does not seek to dictate auditing rules; instead, it aims to identify what aspects require attention, and how the results can contribute to the development and application of reliable AI tools. By emphasising a systemic approach that considers both the technical and social dimensions, our proposed definition of equity serves as a guiding value to foster ethical AI practices.

4 Conclusions

As AI systems become integral to decision-making processes, ensuring their ethical use extends beyond mere technological safety. Dignum [18] discusses the significance of understanding how these systems are designed, the purpose behind their design, and the identity of their designers. Equity adds another layer to ethical AI considerations, addressing the crucial question of who these AI systems are designed for. While, in theory, AI systems should serve the entire society, the reality is that the notion of *all* often lacks the inclusion and diversity it is assumed to possess.

Setting aside discussions of potential existential threats, we think current AI technology has the potential to address some societal issues, but proper accountability measures are imperative to not falling into the trap of technical solutionism [36]. Dignum [18] identifies accountability for results and guiding principles of interactions as foundational for an ethical approach to AI. Regardless of what principle the designers of AI systems choose to use as the basis of their ethics, that principle should be universally understandable and operationalizable across various sectors within multidisciplinary projects. In this paper, we propose equity.

Ethical considerations should be embedded *by-design* within AI systems, but true impact assessment can only be performed when these systems are deployed in society. Real cases, such as Amazon's biased resume screening system [17], reinforce the importance of vigilant monitoring and auditing to intervene before significant harm occurs.

While national and international regulations are on the horizon, AI systems are already in operation. Auditing and monitoring protocols are in place, but their effectiveness varies. This paper proposes defining equity as *providing the necessary resources to those who need them for those people to achieve a sense of belonging within a particular community*, drawing inspiration from feminist thinking. This proposition addresses two prominent challenges in the AI landscape: operationalizing ethical principles in audits, and effectively implementing these audits. Importantly, this definition calls for a social-technical intervention that aligns with the inherently social-technical nature of the AI ecosystem. We hope that our definition will not be seen as a static construct, but rather as an evolving concept ready to withstand scrutiny from at least as diverse a range of expertises as the ones that shaped our definition.

Acknowledgments

Work partly supported by the European Union's Horizon Europe research and innovation programme for the project FINDHR (g.a. No. 101070212). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Holistic AI. 2024. The State of Global AI Regulations in 2024. https://www.holisticai.com/papers/the-state-of-ai-regulations-in-2024 Accessed on March 25, 2024.
[2] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. 3, CSCW (2019), 1–30.
[3] Jose M Alvarez, Alejandra Bringas Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougan, Ioanna Papageorgiou, Paula Reyero, et al. 2024. Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology* 26, 2 (2024), 31.
[4] Nicomachean Ethics1162a Aristotle and VIII Book. 1934. Aristotle in 23 Volumes, Vol. 19, translated by H. Rackham.
[5] Payal Arora and Usha Raman. 2024. The Case for Femwork: Feminist Design Principles in Tech and Tech-Enabled Work. *Computer* 57, 1 (2024).
[6] Simone de Beauvoir. [n. d.]. Les bouches inutiles: pièce en 2 actes et 8 tableaux,[Paris, Théâtre des Carrefours, novembre 1945]. *(No Title)* ([n. d.]).
[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
[8] UN AI Advisory Body. 2023. Interim report: Governing AI for Humanity. https://www.un.org/en/ai-advisory-body Accessed on March 25, 2024.
[9] Philip Brey and Brandt Dainow. 2023. Ethics by design for artificial intelligence. *AI and Ethics* (2023), 1–13.
[10] Alessandra Calvi and Dimitris Kotzinos. 2023. Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1245.
[11] Jeremy Chang, Michael Alan; Roschelle and iSAT Research Group. 2023. Ethics and Equity in AI for Collaborative Learning. https://cacm.acm.org/blogs/blog-cacm/276106-ethics-and-equity-in-ai-for-collaborative-learning/fulltext Accessed on March 25, 2024.
[12] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 598–609.
[13] Mark Coeckelbergh. 2019. Artificial intelligence: some ethical issues and regulatory challenges. *Technology and regulation* 2019 (2019), 31–34.
[14] European Comission. 2024. Artificial Intelligence – Questions and Answers. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683 Accessed on March 25, 2024.
[15] European Commission. 2024. AI history timeline. https://ai-watch.ec.europa.eu/tools/ai-history-timeline_en Accessed on March 25, 2024.
[16] New York City Council. 2023. Law n 2021/144. A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9 Accessed on March 25, 2024.
[17] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*. Auerbach Publications.
[18] Virginia Dignum. 2023. Responsible Artificial Intelligence—From Principles to Practice: A Keynote at TheWebConf 2022. In *ACM SIGIR Forum*, Vol. 56. ACM New York, NY, USA, 1–6.
[19] Luciano Floridi. 2018. Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180081.
[20] Center for AI and Digital Policy. 2024. AI policy frameworks. https://www.caidp.org/resources/ai-policy-frameworks/ Accessed on March 25, 2024.
[21] European Institute for Gender Equality. 2024. What is gender mainstreaming? https://eige.europa.eu/gender-mainstreaming/what-is-gender-mainstreaming?language_content_entity=en Accessed on March 25, 2024.
[22] Paulo Freire. 2019. *Direitos humanos e educação libertadora: gestão democrática da educação pública na cidade de São Paulo.* Editora Paz e Terra.
[23] Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines* 30, 1 (2020), 99–120.
[24] Donna Haraway. 2013. A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. Routledge.
[25] Donna Haraway. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*. Routledge, 455–472.
[26] Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. 2022. Algorithmic bias and risk assessments: Lessons from practice. *Digital Society* 1, 2 (2022), 14.
[27] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[28] David Leslie, Christopher Burr, Mhairi Aitken, jobinsh Cowls, Michael Katell, and Morgan Briggs. 2021. Artificial intelligence, human rights, democracy, and the rule of law: a primer. *arXiv preprint arXiv:2104.04147* (2021).
[29] Audre Lorde. 2012. *Sister outsider: Essays and speeches.* Crossing Press.
[30] Martha Minow. 2021. Equality vs. equity. *American Journal of Law and Equality* 1 (2021), 167–193.
[31] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. 2022. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines* 32, 2 (2022), 241–268.
[32] Jakob Mökander and Luciano Floridi. 2021. Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31, 2 (2021), 323–327.
[33] NESTA. 2021. Participatory AI for humanitarian innovation: a briefing paper. https://www.nesta.org.uk/report/participatory-ai-humanitarian-innovation-briefing-paper/ Accessed on March 25, 2024.
[34] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.
[35] PAI. 2024. The Partnership on AI. https://www.partnershiponai.org/about/ Accessed on March 25, 2024.
[36] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A human rights-based approach to responsible AI. *arXiv preprint arXiv:2210.02667* (2022).
[37] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, et al. 2023. Queer in AI: a case study in community-led participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1882–1895.
[38] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.
[39] John Rawls. 1971. A theory ofjustice.
[40] Salvatore Ruggieri, Jose M Alvarez, Andrea Pugnana, Franco Turini, et al. 2023. Can we trust fair-AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15421–15430.
[41] Giovanni Sartor. 2020. Artificial intelligence and human rights: Between law and ethics. *Maastricht Journal of European and Comparative Law* 27, 6 (2020), 705–719.
[42] Bernd Carsten Stahl. 2021. *Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies.* Springer Nature.
[43] Catharine Titi. 2021. *The Function of Equity in International Law.* Oxford University Press.
[44] Sophie Toupin. 2024. Shaping feminist artificial intelligence. *New Media & Society* 26, 1 (2024), 580–595.
[45] UNESCO. 2023. Guidance on generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693 Accessed on March 25, 2024.
[46] European Union. 2012. Charter for fundamental rights of the European Union. OfficialJournaloftheEuropeanUnion C 326/391.
[47] European Union. 2018. Resolution 2017/3016 (RSP) on Empowering women and girls through the digital sector. Adoptedon17April2018 Strasbourg.
[48] Niels van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170 (2023), 102954.
[49] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 1–10.
[50] Judy Wajcman. 2009. *Technofeminism.* Polity.