Fairness of AI Systems in the Legal Context

Veronica Paternolli[1], Mila Dalla Preda[1,2], and Roberto Giacobazzi[2]

[1] University of Verona, veronica.paternolli@univr.it
[2] University of Verona, mila.dallapreda@univr.it
[3] The University of Arizona, giacobazzi@cs.arizona.edu

Abstract. The digital age has profoundly reshaped societal interactions, heavily influenced by algorithms and Artificial Intelligence (AI). This evolution introduces new challenges in understanding and addressing discrimination, which now arises from both human biases and algorithmic biases, that may cause discriminatory decisions, leading to a form of algorithmic technocracy. AI systems ensure *fairness* when they operate without discrimination. Legal frameworks must adapt to these changes, integrating traditional principles with contemporary technological realities. This paper explores the concept of fairness in AI systems, highlighting the need for both regulatory and technical measures to ensure non-discriminatory practices and to evaluate the accountability for discriminatory behaviors. We present and discuss most commonly used standard mathematical measures for demonstrating fairness, and emphasize the requirements a measure must meet to comply with regulatory aspects of fairness. Our investigation highlights the importance of aligning legal and mathematical approaches to achieve fairness and accountability in AI. We advocate for ongoing assessment and adjustment to maintain ethical standards.

Keywords: Fairness · Mathematical Measures · Legal Codification

1 Introduction

Countries globally are undergoing a profound technological transformation known as the *digital age*, where algorithms and Artificial Intelligence (AI) play a pivotal role in shaping extensive relationships and interactions. This shift has introduced new complexities, particularly in the understanding of discriminatory events. Traditionally viewed as stemming from irrational human behaviour in daily and professional activities, algorithmic decision-making increasingly influences discrimination. Therefore, discrimination must be viewed through the lens of both human and digital influences, creating a novel interplay between human and algorithmic irrationality. Consequently, legal systems, traditionally grounded in human rationality, must undergo significant changes to adapt to these evolving social structures. Thus, as a social science, law is compelled to reassess its core principles in response to the transformative impact of new technologies and the dynamic nature of modern information [29]. Although not all instances of algorithmic discrimination are considered illegal under non-discrimination law, they invariably involve ethical inequities. Consequently, the development of technical measures to identify and attribute responsibility for the perpetuation of discriminatory practices remains a paramount objective within the domain of eXplainable AI (XAI) [2]. Indeed, the recently adopted AI Act [17] requires *fairness* of AI systems, meaning that these systems have to be non-discriminatory. To achieve this, researchers and developers have recurred to established mathematical and theoretical measures to evaluate and demonstrate in courts the compliance of Artificial Intelligence Systems (AIS) with the AI Act [34]. Designing a fair algorithm involves two closely interconnected aspects: first, interpreting and formalizing the requirement of being non-discriminatory within the specific social context; and second, determining the appropriate measures to demonstrate fairness in that context. We agree with Calvi et al. [7] that ensuring fair algorithms requires controllers to regularly assess whether the algorithms are functioning as intended and to adjust them to mitigate biases that may emerge over time.

Moreover, it is crucial to understand the social values and perceptions together with the democratic principles linked to the current mathematical formalization of fairness [23]. For example, consider a facial recognition system implemented in public transport to ensure safety. If the system is designed without adequately considering demographic diversity, particularly for ethnic minorities, it may cause travel delays for these groups and foster a sense of exclusion and discrimination. This scenario might be more pronounced in Italy, where there are stronger privacy protections and a stricter anti-discrimination framework, than in Japan, where the effects and perceptions could differ. A facial recognition system trained predominantly on Japanese faces might achieve high accuracy rates for the majority of users and may not be seen as discriminatory under Japanese laws or societal norms. Additionally, there is greater cultural acceptance of surveillance in Japan for security and public order purposes with respect to Italy.

This paper provides a first step towards the definition of a framework for AI professionals, aiming to bridge the gap between legal requirements and interpretations and scientific assessment methodologies or tools. Specifically, it starts to trace the journey from the legal analysis of the concept of fairness from a European perspective to the examination of mathematical measures for assessing this concept. This offers to technical experts a suitable legal approach to understand and evaluate the effectiveness of the most commonly used measures within their context of use. This is needed because, while the concept of fairness is already qualitatively defined [25,33,40], it remains challenging to apply it quantitatively when addressing emerging social needs in the AI field [22]. The purpose of this work is to facilitate a dialogue between scientific and legal experts in order to establishes the legal boundaries within which computer scientists can navigate in the design and implementation of fair AIS. This holistic approach is fundamental for professionals in the field to develop algorithms that comply with legal standards, thus ensuring that AI systems are both fair and socially responsible.

Outline of the paper: In the second section of the paper, we will analyze the European approach to anti-discrimination legislation, considering the main pillars established by the principal European sources in the matter, including the GDPR and the recent AI Act. Subsequently, in the third section, we will describe some different formalizations of fairness, and propose an analysis of the main mathematical measures used to identify unfair states produced by AI systems and their legal interpretation. Finally, we will conclude by discussing future challenges in the attribution of accountability. We will explore the ongoing issues in defining and applying the principle of fairness in AI systems, and consider potential solutions and directions for future research.

2 Fairness of AIS in the European Union

The segregation caused by certain algorithms used in AIS is under the scrutiny of scientific, legal, and ethical spheres. The investigation focuses on identifying moments of algorithmic failure, the individuals affected by such failures, the resulting social impact, and the entity responsible of the failure [36]. AI technologies already in the market are displaying both intentional and unintentional biases [20]. For instance, an intentional bias can be seen in recruitment algorithms configured to prefer male candidates over female ones for technical positions [24]. On the other hand, unintentional bias is evident in facial recognition systems that perform poorly in identifying individuals with darker skin tones due to imbalanced training datasets. Efforts are directed towards assessing the impartiality, or conversely, the predisposition to bias in algorithmic models and strategies to mitigate such bias when it occurs.

Formally defining what fairness is and what constitutes discrimination is a hard task, as witnessed by the numerous legislative interventions of the European Union (EU) in the recent years. The EU upholds the principle that all individuals are equal before the law and prohibits discrimination on various grounds considering protected characteristics such as gender, race, religion, and disability (as stated by Article 21 of the EU Charter of Fundamental Rights of the European Union (CFREU) [38]). In 2019, this led to the issuance of the Ethics Guidelines for Trustworthy AI, where fairness is identified as one of the requirements that an AIS should meet to be considered trustworthy [3]. Moreover, Article 14 of the European Convention on Human Rights (ECHR) provides a constitutional framework for member states to establish laws to combat discrimination. These laws should protect categories such as sex, race, color, language, religion, political or other opinions, nationality or social origin, association with a national minority, property, birth, or other status [14].

In the EU, fairness typically pertains to non-discrimination, which can be categorized as *direct*, *indirect*, and *intersectional*.

Direct discrimination happens when someone is treated less favorably than another person in a similar situation solely because of their membership in a protected group. The discriminatory action must be directly influenced by the protected characteristic or explicitly taken into account by the decision-maker, with a specific emphasis on the individual affected. Let's consider a facial recognition system used in airport security checks and assume that it shows significantly lower accuracy in recognizing faces of African descent compared to those of European descent. If this system leads to more frequent stops and additional checks for individuals of African descent, it constitutes a clear example of direct discrimination based on ethnic origin.

In contrast, indirect discrimination arises when a policy, criterion, or practice that appears neutral disproportionately disadvantages individuals from a protected group compared to others. For example, consider a company implementing a facial recognition authentication system for access to its buildings. If this system is predominantly trained on male and Caucasian faces, it may result in higher error rates for women and individuals of non-Caucasian origin. Consequently, members of these groups may be required to undergo alternative authentication procedures more frequently, despite the ostensibly neutral nature of the policy. Such discrimination is unlawful unless the policy, criterion, or practice can be objectively justified by a legitimate aim, and the methods used to achieve that aim are both appropriate and necessary.

On the other hand, intersectional discrimination occurs when discrimination involves an individual belonging to multiple protected groups, each of which faces prevalent discrimination [42]. Consider for example a surveillance system with facial recognition in banks that is less accurate in recognising women of Asian origin than Caucasian men or Caucasian women. This leads to a higher number of false positives for Asian women, who are stopped and questioned more frequently. This represents a case of intersectional discrimination, where the interaction between gender and ethnicity creates a specific disadvantage.

Fairness and GDPR: For EU data protection law [16], fairness is a core principle of personal data processing (Art.5(1)(a) GDPR) informing the relationship between data controllers, determining the purposes and means of the processing of personal data, and data subjects, namely the owners of the processed data.

Specifically the GDPR states that the data controller must "*implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests*" when automated decision-making is based on explicit consent or is necessary for the conclusion or performance of a contract. Consequently, it has been convincingly argued that bias minimization strategies must be part of these safeguards. Thus, the absence of strategies for detecting and minimizing bias in this context should be considered a violation of the GDPR (Article 22(3)).

In recent years, the fairness requirements established by the GDPR have enhanced social awareness and perception of the importance of non-discriminatory behaviors [20].

Simultaneously, the GDPR restricts the use of AIS. Specifically, while the GDPR governs the processing of personal data through automated methods, Article 22, explicitly forbids fully autonomous AIS from processing personal data in a way that results in legal consequences for individuals. Consequently, the GDPR mandates that AIS must operate under some form of significant human oversight. Nonetheless, there is a concern that the intricate nature of AI systems could be exploited as a pretext to circumvent rigorous evaluations of AI system outputs for GDPR compliance, where outcomes are superficially endorsed by humans to create the appearance of human oversight and risk assessment.

Moreover, concerning AI discrimination, the GDPR prohibits the processing of special categories of personal data solely by automated means, offering tangible protection against AI discrimination. However, these special categories, as outlined in Article 9 of the GDPR, regrettably do not encompass attributes such as color, language, membership of a national minority, property, and birth, which are referenced in [38]. This exclusion highlights a potential loophole in preventing discriminatory outcomes through personal data processing, whether by AI systems or traditional methods [15].

Fairness and AI Act: The aim of the Regulation by the European Parliament and Council establishing harmonized rules on Artificial Intelligence (referred to as the AI Act) is to enhance the efficiency of the internal market. It aims to achieve this by establishing a consistent legal framework for the development, introduction to the market, deployment, and use of AI systems within the EU. This is aligned with EU values to encourage the adoption of human-centric and reliable AIS, ensuring high levels of health, safety, and fundamental rights protection. Additionally, the AI-Act seeks to mitigate the potential adverse impacts of AIS in the EU while fostering innovation [17]. To this end the AI-Act defines a classification of trustworthy AIS using a risk-based approach [2], particularly concerning biometric identification and categorization of individuals. Specifically the Recital 94 states that: "*Any processing of biometric data involved in the use of AI systems for biometric identification for the purpose of law enforcement needs to comply with Article 10 of Directive (EU) 2016/680, that allows such processing only where strictly necessary, subject to appropriate safeguards for the rights and freedoms of the data subject, and where authorised by Union or Member State law. Such use, when authorised, also needs to respect the principles laid down in Article 4 (1) of Directive (EU) 2016/680 including lawfulness, fairness and transparency, purpose limitation, accuracy and storage limitation*".

However, the AI-Act largely fails to address the identification of root causes and the proposal of solutions to mitigate potential discriminatory impacts caused by AIS. It primarily emphasizes biases in the data sets while neglecting other types of causes, such as those arising from algorithm selection, optimization, or evaluation of mathematical measures. Specifically, biases can be inherent in the data sets used for training, validation, and testing the AI systems and they are often a reflection of historical data patterns or can be introduced during the implementation of AI systems in real-world settings.

The principle of fairness, although not specifically defined in the AI Act, is always accompanied by the requirement of an assessment to counteract discriminatory states that may be produced by the AIS. Based on our analysis, this principle should be grounded in a "*Right to Know All Implications*", which includes understanding all potential fairness violations that may occur when operating the system. This approach allows us to move from the well-studied principle of technical transparency [18], used to assign the risk classification to the AIS, to what we call *implication-transparency*, which encompasses the numerous facets of fairness violation and accountability, and technical transparency.

What has been discussed so far pertains to a qualitative definition of fairness. To concretely identify fairness violations or assess accountability, it is necessary to use mathematical measures in addition to legal categorization. Subsequently, it is essential to return to the legal sphere to interpret the results, as the concept of fairness, from an ethical and legal standpoint, is not fully measurable mathematically but also depends on various human circumstances.

3 Assessing Fairness in AIS

3.1 Formalizing Fairness

In order to establish if an AIS satisfies the fairness requirements it is necessary to provide a more formal definition of fairness together with systematic methodologies for proving it. This has led to different formalizations of the notion of fairness:

- *Group Fairness* ensures that individuals in protected groups receive, on average, the same treatment or outcomes as the overall population [28].
- *Individual Fairness* focuses on guaranteeing that any two individuals who are similar except for protected attributes receive equal or similar treatment or outcomes [19].
- *Causality-Based Fairness* requires that protected attributes, such as gender or race, have no causal effect on outcomes [31].

While it would be ideal to satisfy multiple fairness criteria to achieve comprehensive fairness, this may not be possible due to inherent incompatibilities between the different fairness definitions [4]. Thus, in general, researchers choose the formalization of fairness that better fits the phenomena under analysis, the social context and the legal barrier requirements. Selecting the most appropriate fairness formalization for the context under analysis is a delicate and crucial aspect when assessing the fairness of AIS, that requires both technical and legal expertise. Group Fairness is one of the most common formalizations used in assessing fairness of AIS, as for example in [21]. For this reason, in this work, we focus particularly on Group Fairness, emphasizing how to ensure that AIS treat different demographic groups equitably.

Once we have identified the formal definition of fairness that best fits the context under analysis, it is crucial to identify a suitable measure for proving it. Specifically, it is crucial to understand what needs to be considered in the implementation of a mathematical measure to ensure that the result is functional for assessing the discriminatory state in a given context.

3.2 Fairness Measures

In the following paragraphs, we describe the most commonly used mathematical measures for Group Fairness (such as Equalized Odds, Statistical Parity and

Equal Opportunity), and apply them to the same hypothetical example so to highlight their different implications from a legal point of view. As the use of AI profiling and automated decision-making spreads in the public and private sectors, algorithmic groups are poised to face increasing inequality [41]. This growing concern underscores the importance of robust fairness measures and legal frameworks, which will be further discussed in the following sections of the paper.

Equalizes Odds An example of a discriminatory event in the AI Computer Vision sector could be facial recognition operating with gender and racial bias [9]. In [35] are argued related differents scenarios, such as the significantly poorer performance of commercial facial analysis algorithms, particularly for tasks like gender or smile detection, on images of dark-skinned women is highlighted. These images represent only 7.4% and 4.4% of the widely used benchmark datasets Adience and IJB-A, respectively. As a result, the benchmarking processes on these datasets did not detect or penalize the algorithms' underperformance on this segment of the population [5]. Moreover, in [32] the authors demonstrate that standard PCA can amplify the reconstruction error in one group compared to another one of the same size, as there is no fair method for generating representations with comparable richness across different populations. This makes the dependency on sensitive or protected attributes indistinguishable or hidden [26].

Specifically, this type of discrimination can arise when facial recognition algorithms are trained on unrepresentative datasets [6, 12], primarily containing images of people belonging to a particular ethnic or gender group. As a result, the algorithm may have significantly lower accuracy in recognizing people from other groups. In this context, an appropriate mathematical measure for evaluating and addressing the problem of discrimination is *Equalized Odds*. This measure focuses on equality of opportunity and aims to ensure that the model has comparable performance across different demographic groups. Equalized Odds requires that the probability of obtaining a true positive (True Positive Rate) and a false positive (False Positive Rate) be the same for each protected group. In other words, for any demographic subgroup (e.g., ethnic or gender groups), the algorithm should have similar rates of correct detection and errors [10]. Formally, the Equalised Odds can be defined as follows:

$$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b) \tag{1}$$

$$P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = b) \tag{2}$$

Where $P(M|N)$ denotes, as usual, the conditional probability that the event M will occur given the knowledge that an event N has already occurred and:

- $\hat{Y}$ is the prediction of the model;
- Y is the ground truth;
- A is the protected attribute (e.g. race or gender);
- a and b are different values of the protected attribute.

Table 1: Hypothetical results of facial recognition system

	White Man	Black Women
True positive (TP)	950	800
False positive (FP)	50	200
True Negative (TN)	700	600
False Negative (FN)	300	400

Thus, Equalized Odds is satisfied when the conditional probability of the prediction does not depend from the values assumed by the protected attributes.

Let us consider an AIS that identifies people belonging to two groups: one group is given by white men (Group A) and the second group is made by black women (Group B) representing the possible values of the protected attribute. Let us consider the hypothetical results of the facial recognition system as reported in Table 1.

To calculate the Equalised Odds, we have to compare the True Positive Rate (TPR) and the False Positive Rate (FPR) of the two groups, where:

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{FP + TN} \tag{3}$$

In the considered example we have that the TPR for Group A is 0.76, while for Group B it is 0.67, and the FPR for Group A is 0.07, while for Group B it is 0.25. This shows a significant disparity between the two groups, indicating that the system is less accurate in correctly recognising black women than white men. Hence, according to Equalized Odds, the considered AIS is *not fair*. In order to achieve fairness with respect to Equalised Odds, we should make changes to the system so that the TPR and FPR for both groups are more similar.

Demographic Parity In addition to Equalized Odds, we consider the *Demographic Parity* measure, also known as *Statistical Parity* to illustrate how different fairness measures can provide varying insights depending on the context. Demographic Parity focuses on the equality of positive outcome probabilities across protected groups, regardless of the ground truth, highlighting potential systematic disparities that might not be captured by Equalized Odds. This comparison underscores the importance of selecting the appropriate fairness measure based on the specific requirements of the application.
Demographic Parity requires that the probability of a positive outcome is the same for all protected groups, regardless of the ground truth. The formal definition of Demographic Parity can be expressed as follows:

$$P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b) \tag{4}$$

where:

- $\hat{Y}$ is the prediction of the model;
- A is the protected attribute (e.g., race or gender);
- a and b are different values of the protected attribute.

Let us consider the AIS defined above that identifies people being either white men (Group A) or black women (Group B), whose hypothetical classification results are reported in Table 1.

We consider the probability of a positive outcome for both groups to calculate the Demographic Parity. In the following we use the subscript A to identify the group to which we are referring, so for example TP_A refers to the true positives of group A.

Calculations for Group A:

$$P(\hat{Y} = 1 \mid A = A) = \frac{TP_A + FP_A}{TP_A + FP_A + TN_A + FN_A} = 0.5 \tag{5}$$

Calculations for Group B:

$$P(\hat{Y} = 1 \mid A = B) = \frac{TP_B + FP_B}{TP_B + FP_B + TN_B + FN_B} = 0.5 \tag{6}$$

In this example, Demographic Parity is satisfied because the probability of obtaining a positive outcome is the same for both groups (0.50). This means that, according to the Demographic Parity measure, the considered AIS is *fair*.

Equal Opportunity Ensuring Equal Opportunity in Computer Vision systems is crucial to avoid discrimination and ensure that all people, regardless of their demographic group, have the same opportunity to be correctly recognised when authorised. Specifically, Equal Opportunity requires the true positive rate (TPR) to be the same for all demographic groups, such as white men (Group A) and black women (Group B), as assumed in the two previous examples. Assuming the same values as indicated in Table 1, the TPR for Group A is 0.76 and the TPR for Group B is 0.67.

As highlighted for Equalized Odds, the comparison of the TPR values reveals a significant disparity between the two groups. The TPR for Group A is higher compared to the TPR for Group B. This indicates that the system is more accurate in correctly identifying positive cases for white men than for black women. Thus, according to the Equal Opportunity measure, the AIS is *not fair*.

3.3 Legal Criteria for Choosing Fairness Measures

Demographic Parity may be more appropriate in contexts where ensuring equal opportunities for positive outcomes is critical, while Equalized Odds is more suitable for ensuring fairness in both positive and negative outcomes by considering the ground truth. However, in situations where Equal Opportunity for qualified individuals is the primary concern, the Equal Opportunity measure becomes crucial. This measure focuses on ensuring that all groups have the same chance of receiving a positive outcome when they are qualified for it.

For example, in certain scenarios (like our example reported above), the criterion of fairness according to Equalized Odds may not be met, while Demographic Parity might be satisfied. This indicates a potential issue because Equalized Odds ensures that both the TPR and the FPR are equal across groups. If Equalized Odds is not met, it suggests that the system may have different accuracy and error rates for different groups, leading to unfair treatment.

The system might satisfy Demographic Parity but still be unfair in terms of opportunities. Demographic Parity measures the equality of outcomes, while Equalized Odds measures the equality of opportunity. Satisfying only Demographic Parity does not guarantee that opportunities are distributed equally among groups, thus fully guaranteeing the criterion of Group Fairness on the basis of the prerequisites laid down by law as well. There might be an equal distribution of overall results, but a significant disparity in precision and error rates between different groups. Statistical parity does not respect Group Fairness if the groups have significantly different error rates (TPR and FPR). This means that even if the overall rates of positive outcomes are equal, the experiences and opportunities of the groups can vary significantly, leading to discrimination not detected by Statistical Parity alone. To ensure Group Fairness, it is essential to consider measures like Equalized Odds, which evaluate the equity in error rates across groups.

Considering the context of use, such as in Computer Vision, and the elements of social justice inherent in the protection of individuals who may be discriminated against based on the analysis of their biometric data by automated systems, it becomes evident that the Equalized Odds measure is appropriate. This measure helps highlight violations and supports the assessment activity concerning the European legislation being analyzed, ensuring both detection accuracy and fairness in errors.

Thus, the choice of the appropriate measures for fairness should consider the following aspects:

- the social context related to the type of discrimination perpetrated by the acting subject (AIS);
- the criterion of fairness considered;
- the type of outcomes produced (positive or negative), whether or not connected to the ground truth.

4 Fair Perspectives of Artificial Decision-Making Systems

4.1 Accountability

The development of AI poses a significant challenge to existing liability frameworks. The legislation generally ensures that a person who suffers harm or damage has the right to seek compensation from the party deemed accountable and to receive compensation from that party. On the other hand, it provides economic incentives for individuals to avoid causing harm or damage in the first place. In this context, the level of diligence expected from an AI specialist should be proportionate to i) the nature of the AI system, ii) the legally protected right potentially affected, iii) the potential harm or damage that the AI system could cause, and iv) the likelihood of such harm [27]. To meet this regulatory requirement which has been explained, it is essential to establish a close connection between the concept of accountability and that of fairness [1], and more specifically on the *implication-transparency.* This need also arises to address the demands of the AI Act [39], which operates across fundamental rights and individual freedoms, proposing a human-centered and reliable approach. Thus, the principle of algorithmic non-discrimination must be anchored in comprehensive oversight, including both human control and human-machine collaboration [37].

This can only be achieved through the development of an up-to-date legal model that integrates a clear and current framework on how specific biases are introduced and propagated in the numerical implementation of the algorithm itself, which does not currently exist. The legal ecosystem lacks a clear framework for attributing accountability when unjustified harm is inflicted upon a passive subject and a structural model approach to identify discriminatory causes and the corresponding range of explanations. This is increasingly important as AIS are recognized as gradually intelligent entities, effectively acting subjects.

4.2 AI Open Challenges

Current scientific contributions propose classifications of the concept of fairness through a limited *legal-informatics* perspective [13], focusing the greatest effort on implementing mathematical measures without fully understanding the legal explanation of the results. The concept of explanation, which is the result of combining legal techniques and mathematical measures, should play an important role in the AI & Law community, being related to the general quest for justification and transparency of legal decision-making. Within an argumentation-based approach, the justification of a legal assumption may be viewed as an argument structure aimed to show that the decision is right or correct, according to a convincing reconstruction of practical facts and norms [30].

As illustrated by examples from the field of Computer Vision, integrating legal and technical frameworks has become essential for identifying significant disparities and ensuring that professionals in the sector comply with the legal and ethical guidelines imposed by the ecosystem. Furthermore, beyond the metrics employed in this study, there are numerous others worth exploring to establish guidelines grounded in legal awareness [9]. This approach would facilitate reasoning that identifies the most appropriate metrics based on the relevant legal criteria.

This approach would address the supreme need identified by the embryonic European framework on civil liability (Liability Rules for Artificial Intelligence), which expressly states: "*considering that certain AI systems present significant legal challenges to the current liability framework and could lead to situations where their opacity may make it extremely burdensome or even impossible to* *identify who had control over the risk associated with the AI system or which code, input, or data ultimately caused the harmful activity; this factor could make it more difficult to establish the link between the harm or damage and the behavior that caused it, resulting in victims potentially not receiving adequate compensation*" [8]. Therefore, the greatest ongoing challenge is to create a taxonomy of the different fairness measures used to assess the main notions of fairness (Group Fairness, Individual Fairness, and Causality-Based Fairness) and to identify, for each of these, the most appropriate mathematical measure to meet the legal constraints in a given context of use. This is essential for assigning responsibility for the potential failure of the AIS system and the realization of discriminatory outcomes.

5 Final Discussion

We have begun analyzing the different formalizations of fairness and the most commonly used measures to demonstrate it. However, as highlighted above, further work is needed to extend the analysis of these measures and notions of fairness. The interpretation of fairness remains ambiguous, with each party—whether accusing or defending—interpreting it according to their perspective. For instance, in the cases of Amazon and COMPAS [43, 44], this lack of a unified understanding of fairness has led to considerable controversy. In the second scenario, the probability of a defendant reoffending is estimated, which can assist judges or parole officers in making decisions regarding pre-trial release. Models of this type often rely on proxy variables like "arrest" to represent "crime" or to capture an underlying concept of "riskiness." Due to the increased level of policing in minority communities, these proxies are often misrepresented, leading to a different relationship between "crime" and "arrest" for individuals from these communities. Other variables used in COMPAS, such as "rearrest" as a proxy for "recidivism" [11], also suffer from similar inaccuracies. Consequently, the model produced a significantly higher rate of false positives for Black defendants compared to White defendants, meaning it was more prone to incorrectly assessing Black defendants as high-risk for reoffending when they were not. Therefore, establishing criteria to decide on the formalization of fairness and the measures based on context would help eliminate the ambiguity in individual interpretations. This is a challenging problem, as it involves considering a multitude of factors, including legal, cultural, social, political, and algorithmic aspects. We believe that the only way to effectively address and resolve the legal issues associated with the increasing use of AI systems in any aspect of society is through a dialogue between AI experts and legal professionals a dialogue that forms the foundation of the present work.

Acknowledgement. This work was partially supported by the project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

1. Akinrinola, O., Okoye, C., Ofodile, O., Ugochukwu, C.: Navigating and reviewing ethical dilemmas in ai development: Strategies for transparency, fairness, and accountability. GSC Advanced Research and Reviews **18**, 050–058 (03 2024). https://doi.org/10.30574/gscarr.2024.18.3.0088
2. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion **99**, 101805 (04 2023). https://doi.org/10.1016/j.inffus.2023.101805
3. on Artificial Intelligence, E.C.H.L.E.G.: Ethics guidelines for trustworthy ai (2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419, accessed: 2024-06-28
4. Bringas Colmenarejo, A., Nannini, L., Rieger, A., Scott, K.M., Zhao, X., Patro, G.K., Kasneci, G., Kinder-Kurlanda, K.: Fairness in agreement with european values: An interdisciplinary perspective on ai regulation. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. AIES '22, ACM (Jul 2022). https://doi.org/10.1145/3514094.3534158, http://dx.doi.org/10.1145/3514094.3534158
5. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. pp. 77–91 (2018)
6. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (23–24 Feb 2018), https://proceedings.mlr.press/v81/buolamwini18a.html
7. Calvi, A., Malgieri, G., Kotzinos, D.: The unfair side of privacy enhancing technologies: addressing the trade-offs between pets and fairness. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3630106.3659024, https://doi.org/10.1145/3630106.3659024
8. Commission, E.: Liability rules for artificial intelligence (2022), https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en, accessed: 2024-07-08
9. Dominguez-Catena, I., Paternain, D., Galar, M.: Metrics for dataset demographic bias: A case study on facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5209–5226 (Aug 2024). https://doi.org/10.1109/tpami.2024.3361979, http://dx.doi.org/10.1109/TPAMI.2024.3361979
10. Dominguez-Catena, I., Paternain, D., Galar, M.: Metrics for dataset demographic bias: A case study on facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–18 (2024). https://doi.org/10.1109/tpami.2024.3361979, http://dx.doi.org/10.1109/TPAMI.2024.3361979
11. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science Advances **4**(1), eaao5580 (Jan 17 2018). https://doi.org/10.1126/sciadv.aao5580
12. Dulhanty, C., Wong, A.: Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. CoRR **abs/1905.01347** (2019), http://arxiv.org/abs/1905.01347

13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness (2011)
14. of Europe, C.: Article 21 of the european convention on human rights (echr) (1950), https://www.echr.coe.int/Documents/Convention_ENG.pdf, accessed: 2024-06-28
15. European Papers: Ai regulation through the lens of fundamental rights (2024), https://www.europeanpapers.eu/en/europeanforum/ai-regulation-through-the-lens-of-fundamental-rights, accessed: 2024-06-30
16. European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj (2016), accessed: 2024-07-07
17. European Union, European Parliament, and Council of the European Union: Regulation (EU) 2024/... of the European Parliament and of the Council 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) (2024), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
18. Felzmann, H., Fosch-Villaronga, E., Lutz, C., et al.: Towards transparency by design for artificial intelligence. Science and Engineering Ethics **26**, 3333–3361 (2020). https://doi.org/10.1007/s11948-020-00276-4, https://doi.org/10.1007/s11948-020-00276-4
19. Fleisher, W.: What's fair about individual fairness? Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3461702.3462621, https://doi.org/10.1145/3461702.3462621
20. Hacker, P.: Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. Common Market Law Review **55**, 1143–1186 (2018), available at SSRN: https://ssrn.com/abstract=3164973
21. Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., Hu, X.: Generalized demographic parity for group fairness (02 2022)
22. John-Mathews, J., Cardon, D., Balagué, C.: From reality to world. a critical perspective on ai fairness. Journal of Business Ethics **178**(4), 945–959 (2022). https://doi.org/10.1007/s10551-022-05055-8, https://doi.org/10.1007/s10551-022-05055-8
23. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. AEA Papers and Proceedings **108**, 22–27 (2018)
24. Lapowsky, I.: Google autocomplete still makes vile suggestions. Wired (2018), https://www.wired.com/story/google-autocomplete-still-makes-vile-suggestions/, available at Wired.com
25. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. CoRR **abs/1908.09635** (2019), http://arxiv.org/abs/1908.09635
26. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. CoRR **abs/1908.09635** (2019), http://arxiv.org/abs/1908.09635
27. Parliament, E.: Resolution of 20 october 2020 on the framework of ethical aspects of artificial intelligence, robotics and related technologies (2020), https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_IT.html#title1, accessed: 2024-06-12

28. Räz, T.: Group fairness: Independence revisited. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445876, https://doi.org/10.1145/3442188.3445876
29. Rodrigues, R.: Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. Journal of Responsible Technology **4**, 100005 (10 2020). https://doi.org/10.1016/j.jrt.2020.100005
30. Rotolo, A., Sartor, G.: Argumentation and explanation in the law. Frontiers in Artificial Intelligence **6**, 1130559 (Sep 2023). https://doi.org/10.3389/frai.2023.1130559
31. Rychener, Y., Taşkesen, B., Kuhn, D.: Metrizing fairness (05 2022). https://doi.org/10.48550/arXiv.2205.15049
32. Samadi, S., Tantipongpipat, U.T., Morgenstern, J., Singh, M., Vempala, S.S.: The price of fair PCA: one extra dimension. CoRR **abs/1811.00103** (2018), http://arxiv.org/abs/1811.00103
33. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a standard for identifying and managing bias in artificial intelligence. Special publication (nist sp), National Institute of Standards and Technology, Gaithersburg, MD (2022). https://doi.org/10.6028/NIST.SP.1270, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464, [online], Accessed August 9, 2024
34. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: Metrics, explainability and the european ai act proposal. J **5**, 126–138 (02 2022). https://doi.org/10.3390/j5010010
35. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. CoRR **abs/1901.10002** (2019), http://arxiv.org/abs/1901.10002
36. Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.: The ethics of algorithms: key problems and solutions. AI SOCIETY **37** (03 2022). https://doi.org/10.1007/s00146-021-01154-8
37. Tsamados, A., Floridi, L., Taddeo, M.: Human control of ai systems: from supervision to teaming. AI and Ethics (05 2024). https://doi.org/10.1007/s43681-024-00489-4
38. Union, E.: Article 21 of the eu charter of fundamental rights of the european union (cfreu) (2000), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A120
39. Union, E.: Regulation (eu) 2021/0106 on a european approach for artificial intelligence (2024), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, accessed: 2024-06-12
40. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3194770.3194776, https://doi.org/10.1145/3194770.3194776
41. Wachter, S.: The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. Tulane Law Review **97**, 149 (2022). https://doi.org/10.2139/ssrn.4099100, https://ssrn.com/abstract=4099100
42. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. Computer Law & Security Review **41**, 105567 (2021). https://doi.org/10.1016/j.clsr.2021.105567, https://www.sciencedirect.com/science/article/pii/S0267364921000406
43. Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K.P., Weller, A.: An empirical study on learning fairness metrics for compas data with human supervision (2019), https://arxiv.org/abs/1910.10255

44. Zhao, Y., Zhang, X., Tang, X., Qin, C., Zhu, H.: Embedding fairness into the ai-based talent recruitment systems: The perspective of environment cycle and knowledge cycle. In: PACIS 2021 Proceedings. No. 15 (2021), https://aisel.aisnet.org/pacis2021/15