Computer Law & Security Review 55 (2024) 106053

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

Assessing trustworthy AI: Technical and legal perspectives of fairness in AI

Markus Kattnig [a], Alessa Angerschmid [a], Thomas Reichel [a], Roman Kern [a,b,*]

[a] Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36, Graz, 8010, Styria, Austria
[b] Know Center Research GmbH, Sandgasse 34, Graz, 8010, Styria, Austria

ARTICLE INFO

Keywords:
Fairness
Non-discrimination
Group fairness
Individual fairness
AI Act
Trustworthy AI
Bias

ABSTRACT

Artificial Intelligence systems are used more and more nowadays, from the application of decision support systems to autonomous vehicles. Hence, the widespread use of AI systems in various fields raises concerns about their potential impact on human safety and autonomy, especially regarding fair decision-making. In our research, we primarily concentrate on aspects of non-discrimination, encompassing both group and individual fairness. Therefore, it must be ensured that decisions made by such systems are fair and unbiased. Although there are many different methods for bias mitigation, few of them meet existing legal requirements. Unclear legal frameworks further worsen this problem. To address this issue, this paper investigates current state-of-the-art methods for bias mitigation and contrasts them with the legal requirements, with the scope limited to the European Union and with a particular focus on the AI Act. Moreover, the paper initially examines state-of-the-art approaches to ensure AI fairness, and subsequently, outlines various fairness measures. Challenges of defining fairness and the need for a comprehensive legal methodology to address fairness in AI systems are discussed. The paper contributes to the ongoing discussion on fairness in AI and highlights the importance of meeting legal requirements to ensure fairness and non-discrimination for all data subjects.

1. Introduction

One of the current major research topics in the field of Artificial Intelligence (AI) is ensuring fairness in AI systems [1–4]. In recent years, the growing influence of increasingly skilled AI systems on people's everyday lives has highlighted the importance of fairness. A survey from 2022 shows that many organisations recognise AI as a future technology and desire to invest in it [5]. The increasing shift from human decisions to machine-made ones is an ongoing development due to the increasing digitalisation in our society [6]. While AI systems have shown immense potential to decrease human workload and improve accuracy [7], they also pose a risk to human safety and autonomy [8]. One significant concern hereby is the risk of a bias leading to unjust decisions. Here, bias refers to the presence of systematic and unfair behaviour and errors in AI systems. Hence, a bias may lead to discriminatory outcomes and unfair treatment, which further implicates the importance of fairness. For illustration, in the case of the COMPAS algorithm [9], the prevalence of a racial bias has been found [10]. Such bias is inadvertently introduced into AI systems when data contains systematic or historical inequalities, which exacerbate unfair treatment of already disadvantaged groups, leading to serious social and ethical implications [11]. As a further complication, the identification of a bias is in general challenging. Given the prevalence of AI systems and the potential risks involved, it is crucial to ensure that decisions made by these systems are fair and unbiased. To achieve this, various methods for bias mitigation have been proposed [4,12,13].

However, another major concern is the lack of common sense within AI systems. This concept is also referred to as Causality, the ability to infer cause and effect [14]. This implies that AI systems are not properly able to grieve a concept and apply a solution to a new, unknown problem [15]. It should be kept in mind, that these systems do not "reason" the same way, a human would do.

While this article mainly aligns its concept of fairness with the principles of EU non-discrimination and equality law, it also acknowledges that fairness is a broader concept encompassing a wider range of legal considerations. However, it is an open challenge to meet the existing legal requirements for fairness and non-discrimination, particularly due to ambiguous legal frameworks [16]. Accordingly *"fairness"* itself is a complex concept, and defining it requires an interdisciplinary approach. Especially since the term fairness is not uniformly defined and is often understood differently in scientific disciplines. It is quite perplexing to accurately define fairness using rigorous *"mathematical terms"*. This task proves to be more challenging than one would expect due to the ambiguity surrounding the concept.

Ensuring fairness in AI systems has become a research topic of high interest in recent years in computer science. There is a lack

* Corresponding author at: Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36, Graz, 8010, Styria, Austria.
E-mail address: rkern@tugraz.at (R. Kern).

https://doi.org/10.1016/j.clsr.2024.106053

Available online 18 September 2024

of comprehension of how the concept of fairness from the legal and technical perspectives relate to each other. Although many methods for mitigating bias in AI systems have been developed, few of them meet legal requirements. Focusing on the European area, the European Union (EU) aims to influence the development and use of AI systems by defining legal requirements. Hence, it is crucial to examine the legal requirements for AI systems to ensure fairness and non-discrimination. This paper examines the legal framework for AI in the EU and provides insights into the challenges and opportunities of addressing bias mitigation in AI systems in the EU from an interdisciplinary perspective and contributes to this by examining the legal requirements for AI systems to ensure fairness and non-discrimination.

The paper informs about the importance of ensuring fairness in AI systems and thus the need for an interdisciplinary approach to understanding the legal and ethical principles of fairness. In this context, this paper provides deeper insights into current state-of-the-art methods for bias mitigation in AI systems and contrasts them with the legal requirements of the European Union. In addition, the challenges of defining fairness and the need for a comprehensive legal methodology to approach fairness in AI systems are addressed. Achieving fairness in AI decision-making is not only crucial to avoid discriminatory outcomes and unfair treatment, but it is also essential to promote social justice and ethical values. To this end, it is important to develop and implement effective methods for bias mitigation that meet legal requirements and ensure fairness and non-discrimination in AI systems. This paper concludes with recommendations for establishing standards and driving future developments in fair AI systems across technology, science, law, and policy-making.

2. Fairness definitions

Fairness is considered the central starting point for the application of AI in society [11]. However, fairness is not uniformly defined. The concept of fairness is already subjective in its structure. There is no agreed notion of "fairness" since various scientific communities approach fairness differently. The respective conceptualisations of fairness differ due to various scientific backgrounds in the scientific discourse [17].

On a personal level, fairness depends on the subjective feeling of a person. Nevertheless, there is a common and universal understanding of the term fairness within a society [18]. This is evident when reporting on criminals and their convictions. So, often a majority can be found, who are either satisfied with the punishment and see it as fair or those who are dissatisfied and suspect unfairness. Moreover, the overall notion of fairness tends to be ubiquitous across cultures [19]. Although this intuitive understanding of the concept of fairness exists, it does not follow a stringent definition.

As a consequence, legislators often work with the term without defining it precisely. For example, the European Commission writes that AI should be used "fairly" [20]. This lack of clarity has led to an ongoing scientific discourse. Not only are there numerous mathematical definitions, but also sociological and ethical definitions [21–24]. In addition, for many technical realisations, typically fairness is either based on individuals or groups of people. As a consequence, depending on the understanding of fairness being used, outcomes obtained from AI systems vary [25].

By examining the legal definition of fairness, and the technical aspects related to its implementation, this paper endeavours to present a clear and cohesive understanding of fairness. In doing so, it emphasises the essential similarities between these differing perspectives, which are necessary to establish a unified approach towards achieving a common understanding of fairness.

2.1. Legal fairness

Fairness is repeatedly accentuated in the legal sciences, but it remains a theoretical concept and its implementation in practice is lacking [26]. The primary context for comprehending fairness involves equality and equity. While equality emphasises equal treatment for all individuals, equity emphasises justice in the distribution of resources or opportunities based on individual needs [27]. Overall fairness refers to the principle that legal procedures and processes should be fair and impartial. This is commonly interpreted that all persons are equal in relation to the legal system and treated with dignity, respect, and equality [28].

Moreover, the use of AI systems might lead to the restriction of rights or opportunities of individuals [29]. Hence, automated decision-making, or even decision support systems could be problematic, not only based on their data usage but also due to their potential influences on individuals.

Procedural fairness. The concept of fairness is very closely linked to the broader concept of justice, as shown by Colquitt [30], who conducted a study that showed the multi-dimensional nature of organisational justice. Hereby, organisational justice refers to an individual's perception of events within an organisation to be impartial and fair [31]. This also implies decisions on salaries, projects carried out or social settings within the organisational structure [30]. In contrast, the term procedural justice refers to the exercise of authority by legal authorities [32]. In this sense, Rawls [33] argues that court proceedings should focus on the acceptance of all involved parties, regardless of their beliefs or personal interests. Processes and procedures must therefore be designed in such a way that there is no discrimination, bias or prejudice towards individual parties. Lind and Tyler [34] showed that processes in the court system significantly influence people's image of procedural fairness, which also influences people's compliance with the law. Furthermore, Lind and Tyler [35] identified three main components of procedural justice, which are (1) the opportunity to voice an opinion, (2) the perception of decision-makers as fair and unbiased, and (3) the level of respect and dignity shown to the parties.

However, justice and fairness are two different concepts that cannot be used interchangeably. While justice means compliance with applicable laws, fairness means how people react to the law [36]. The derived question of which elements influence the perception has therefore led to the proposal of various frameworks for understanding and defining fairness. This was further defined by Tyler [28], who identified four key elements of procedural fairness, see Table 1.

Table 1
List of key elements of procedural fairness and their description according to [28].

Key element	Description
Voice	The opportunity to be heard
Neutrality	The impartiality of the decision-maker
Respect	The treatment with dignity and politeness
Trust	The perceived legitimacy of the process

Substantive fairness. In addition to procedural fairness, notable reference should also be made to substantive fairness which refers to the content and outcome of decisions or processes and ensures that they are just and equitable, based on an assessment of the merits and facts of a case [37]. In contrast to procedural fairness, which focuses on the fairness of the process that leads to a decision, substantive fairness focuses on the decision itself. In certain fields such as law, labour, and public policy, it is important that decisions are made based on fair and equitable criteria. This ensures that individuals' rights are respected and unjust outcomes are avoided. In the context of employment, this could involve ensuring that hiring and promotion practices are non-discriminatory and based on an individual's actual merits and qualifications [38]. Additionally, it may involve ensuring that employees are paid fairly for their work. Substantive fairness is sometimes viewed as being similar to procedural fairness. The argument is that a reasonable and knowledgeable person would not accept unfair treatment such as disadvantageous contract terms or prices that exceed the market rate [37]. Rather substantive fairness looks at

whether the actual terms of an agreement or a court decision are fair, not just if the process to reach that agreement or decision was fair. Even if all the rules were followed in a contract negotiation or court case, the final outcome should be fair to everyone involved. This means that the outcome should consider more than just the steps taken to get there. The criteria for the outcome should also be just and take into account the interests of all parties [39].

Legislative realisation. These elements are also reflected in the legal frameworks [40]. In Article 6 of the *"European Convention on Human Rights"* (ECHR) [41] the right to a fair trial is enshrined, which is considered a fundamental pillar of the rule of law and guarantees the separation of powers [42]. Article 6(1) stipulates that any determination of civil rights and obligations or of any criminal charge has to be made within a reasonable time by an independent, and impartial tribunal established by law. In addition, the verdict must be announced publicly. In Paragraphs 2 and 3, further regulations are laid down for criminal offences, which are also applicable by analogy in civil cases [43]. This includes in particular the presumption of innocence in accordance with Art 6(2). Case law has created new guarantees that are not expressly mentioned in Article 6. According to the opinion of the court, these guarantees are based on the underlying concept of fair proceedings [42]. These additional guarantees include access to a court, the right to legal aid and equality of arms.

In this context, reference should be made to the *"Charter of the Fundamental Rights of the European Union"* (CFR) [44], in which Article 47 enshrines the right to an effective remedy and a fair trial. Article 47 CFR is divided into three paragraphs, with the first paragraph stating that everyone, *"whose rights and freedoms guaranteed by the law of the Union are violated, has the right to an effective remedy before a tribunal in compliance with the conditions laid down in this Article"*. According to the second paragraph, *"everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal, previously established by law and shall have the possibility of being advised, defended, and represented"*. Finally, the third paragraph states that *"legal aid shall be made available to those who lack sufficient resources in so far as such aid is necessary to ensure effective access to justice"*. Comparing Article 6 ECHR and Article 47 CFR, a certain similarity is evident. The close relationship between these articles can also be deduced from the fact that the European Court of Justice refers to decisions of the European Court of Human Rights [45]. Overall, this indicates that the legislature primarily considers fairness in the context of court proceedings.

The situation is similar in the area of arbitration. Arbitration proceedings are governed by fewer procedural rules compared to court proceedings, allowing for more flexibility in resolving disputes, and thus providing parties with greater control over the outcome of their case. Moreover, there is no evidence that plaintiffs fare worse in arbitration than in regular courts; in fact, studies show that they do about as equally well [46]. Here, too, fairness is a central aspect of the procedure. The arbitration procedure also attaches central importance to the aspect of fairness, which is also derived from Article 6 of the ECHR. The international documents on the rules of arbitration courts also contain references to fairness in several places. For example, the *"The Code of Ethics for Arbitrators in Commercial Disputes"* of the American Arbitration Association [47], defines in Canon 1 A that *"an arbitrator has a responsibility not only to the parties but also to the process of arbitration itself, and must observe high standards of conduct so that the integrity and fairness of the arbitration are guaranteed. standards of conduct so that the integrity and fairness of the process will be preserved"*.

The *"Rules of Arbitration and Mediation of the Vienna International Arbitral Centre"* (VIAC) [48] also states in Article 28(1) that *"the arbitral tribunal shall conduct the arbitration in accordance with the Vienna Rules and the agreement of the parties in an efficient and cost-effective manner, but otherwise as it deems appropriate. The arbitral tribunal shall treat the parties fairly. The parties shall have the right to be heard at any stage of the proceedings"*. It should be noted that equal treatment of the parties not only includes the principle of equal treatment but also other forms. An example of this is the problem of formal equal treatment, which can be unfair for one party in the end since short deadlines for filing documents can be disadvantageous for the party with the burden of proof. Similarly, in the area of claims for damages based on errors in AI systems, it is often difficult for the plaintiff to obtain information on the structure and function of the AI system in a short time and subsequently evaluate this information in the appropriate time to substantiate the claim.

In addition, a large number of other fairness concepts can be found in the various pieces of legislation. For example, the GDPR, which sets out such a fairness principle in Article 5(1)(a) in particular, which assumes that personal data is processed *"fairly"* if personal data is processed in compliance with the GDPR.

Likewise, fairness is also incorporated in the EU competition law. This is clarified by Article 102 TFEU, where it is stated that *"Any abuse by one or more undertakings of a dominant position within the internal market or in a substantial part of it shall be prohibited as incompatible with the internal market in so far as it may affect trade between Member States"*. On this note, the Data Act [49] sets out the objective of creating *"fair"* access to and use of data as well as *"fair"* and competitive market conditions with regard to the rules for sharing the data generated. Given this context, Article 13 is particularly worth noting. It introduces a pivotal shift in contractual dynamics by instituting a criterion of fairness for specific provisions within data agreements. This article stipulates that a contractual clause shall not be binding if it is deemed unfair. Furthermore, Article 13 delineates a comprehensive fairness evaluation framework, which includes a broad unfairness assessment guideline under Section 3, according to which a clause is unfair *"if it is of such a nature that its use grossly deviates from good commercial practice in data access and use, contrary to good faith and fair dealing"*.

A similar fairness principle can be found in the EU consumer protection law, such as the Unfair Terms Directive (UTD) [50] or the Unfair Commercial Practices Directive (UCPD) [51]. The UTD focuses on protecting consumers from unfair terms in contracts they have not individually negotiated, ensuring contract fairness and transparency, whereas the UCPD safeguards consumers against misleading and aggressive business practices, promoting honest market conduct and informed decision-making.

AI-related fairness. With the advent of Artificial Intelligence and the steadily growing areas of its application, the focus has expanded from fair trials to ensuring fairness in AI systems. In general, an AI system is *"a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment"* [52]. Given this background, AI in this paper is to be understood according to the definition of AI-system in the AI Act [53], where it is defined in Article 3(1) as *"a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"*. Not only must the concept of fairness be adapted to the needs of AI [54], it must also be ensured that the use of AI does not weaken the right to a fair trial [55].

Moreover, it should be noted that "unfair" AI systems might be created due to an unfitting design or the inclusion of inappropriate data [29]. Thus, introducing the problem of algorithmic transparency, which can lead to unjust decisions. Of course, this could in turn access the opportunities of individuals, if the decisions are treated as facts.

In this sense fairness in relation to AI primarily refers to the aspect of non-discrimination [56–58]. In this sense, fairness is very closely related to the idea that every person is treated equally. Ensuring non-discrimination is one of the fundamental ideas of the European Union

and was laid down in the *"Treaty on the Functioning of the European Union"* (TFEU) [59]. Reference should be made to Article 10 TFEU, where it stated that *"the Union shall aim to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation"*, and also to Article 19 TFEU, where it is clearly emphasised that the *"appropriate action to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation"* should be pursued by the European Council.

A main source of law hereby is Article 21 CFR. According to this, it is legislated that *"any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation"* is prohibited. Hence, fairness is the absence of unfairness. The European legislator has therefore issued essential provisions which prohibit unequal treatment based on given characteristics. Further regard should be made to the *"Race Equality Directive"* [60] and the *"Framework Equality Directive"* [61].

A fundamental distinction is to be made between direct and indirect discrimination. Direct discrimination occurs when a person or a group is treated less favourably because of a specific characteristic such as gender or race. Indirect discrimination occurs when an apparently neutral provision puts persons of specific groups at a particular disadvantage compared with other persons. Direct discrimination can never be justified and is therefore prohibited in any case, while indirect discrimination can be objectively justified by a legitimate aim if the means of achieving that aim are appropriate and necessary [62]. Typically discrimination by AI also implies the term bias. A bias describes a tendency that prevents an unprejudiced decision. Such a bias can exist in people due to prejudices. The term bias is used in the technical field to describe the problem of biased decision-making [56]. Research is concerned with ensuring unbiased decision-making and thereby preventing discrimination by AI systems.

With regard to AI systems, it should also be emphasised that these can inadvertently classify people into disadvantaged groups based on apparently unrelated features. This correlation leading to discrimination is also called proxy discrimination [63]. Hence, in the case of AI systems, it is often not direct discrimination being the most problematic, but indirect discrimination, since this is typically hard to detect [64]. The danger that AI poses in this regard has been recognised by the European Union. For example, the Gender Equality Strategy 2020-2025 [65] expressly states that *"algorithms and related machine-learning, if not transparent and robust enough, risk repeating, amplifying or contributing to gender biases that programmers may not be aware of or that are the result of specific data selection"*. The problem of discrimination is also addressed in the draft of the new *"AI Act"* [66]. In this regard Recital 36 states that *"AI systems used for this purpose may lead to discrimination of persons or groups and perpetuate historical patterns of discrimination, for example, based on racial or ethnic origins, disabilities, age, sexual orientation, or create new forms of discriminatory impacts"*.

Furthermore, the version of the AI Act of 14 June 2023 [67] also fails to adequately address the concept of fairness. In the amended version Article 4a(1)(e) states that *"diversity, non-discrimination and fairness' means that AI systems shall be developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law"*. Particularly striking is the blending of several concepts into a single term, which does not seem appropriate. Although there is a certain affinity between the terms diversity, non-discrimination and fairness, they are associated with other different aspects which are not identical. Mixing them in a single definition is therefore not sufficient to clarify the legislative intention. Even with special attention to the aspect of fairness, it cannot be deduced under which circumstances fairness is actually satisfied. In this context, the question also arises as to when an *"unfair"* bias is to be assumed. Rather, it seems necessary to define them separately to create clarity. Consequently, this legal definition does not effectively address the persistent confusion surrounding these terms.

The final version of the AI Act [53] does not appear to contribute to a clearer definition of fairness. It should be noted that the final text of the regulation does not contain a comprehensive thematisation of fairness. Compared to the previous version, it should be noted that fairness is given even less importance. For example, the previously cited Article 4a(1)(e) is no longer included in the AI Act, and references to fairness can certainly be found primarily in the recitals, specifically in Recital 27, 48, 94 and 110. In addition, indirect references can be found in Recital 59, 61 and 74, whereby the right to an effective remedy and to a fair trial is referred to principally. The fact that fairness is no longer specifically addressed in an article, and rather in the recitals, leads to a further dilution of its importance. Recitals in EU law provide context and explanations for the regulations and directives but are not legally binding themselves, whereas articles within these legal instruments are operative provisions that lay down legal obligations and rights. Only articles have a binding character and are therefore enforceable. This fundamental difference makes articles more important for legal compliance and enforcement because they directly impact the legal duties and rights of entities and individuals under EU law. Given this background, the AI Act fails to recognise the regulative need for a concept of fairness and its corresponding elements. Only the reference to the Ethics guideline for trustworthy AI, which is made in Recital 27, according to which the principles mentioned in the guideline, which include in particular the principle of fairness, should be *"translated, when possible, in the design and use of AI models"*, focuses on this topic. However, this wording alone represents an inadequate observance of fairness, as it should only be considered, *"when possible"*. Similarly, Recital 165 also only expresses in relation to high-risk AI systems that *"Providers and, as appropriate, deployers of all AI systems, high-risk or not, and AI models should also be encouraged to apply on a voluntary basis additional requirements related, for example, to the elements of the Union's Ethics Guidelines for Trustworthy AI"*.

Respectively, particular criticism is levelled at the fact that under current EU law, grounds of discrimination are considered individually instead of as a whole, as would be the case with the concept of intersectional discrimination. This becomes clear by focusing on the case Parris (2016) [68], in which the ECJ ruled that *"there is, however, no new category of discrimination resulting from the combination of more than one of those grounds, such as sexual orientation and age, that may be found to exist where discrimination on the basis of those grounds taken in isolation has not been established"*. Additionally, traditional non-discrimination law primarily covers explicitly protected characteristics such as race and gender. However, they may fall short in countering the subtler forms of discrimination that AI can introduce or amplify, particularly concerning non-traditional sensitive attributes like economic or social status. Proxy discrimination emerges when AI utilises neutral factors as stand-ins for these protected attributes, inadvertently leading to biased outcomes. This highlights the urgent need for revising non-discrimination frameworks and advocating for ethical AI practices. Hence, in jurisprudence, an ongoing discussion exists as to whether the current non-discrimination rules are a useful tool for regulating AI at all or whether significant adjustments are required [57,69,70].

Although nowadays the focus is no longer just on court proceedings, due to the increasing use of AI systems, the four elements of procedural fairness established by Tyler [28] still hold significant importance to ensure fairness in AI systems. Focusing on fairness only in relation to non-discrimination is by no means sufficient to meet the requirements of trust and transparency. Non-discrimination can be assigned to the element of respect since any person should be treated with dignity and politeness [71]. However, the elements of voice, neutrality and trust remain open as well as the core elements of substantive fairness.

2.2. Technical fairness

In the technical field, the term fairness is defined more clearly, though there is no commonly agreed "best" definition of fairness [2,72,73]. Hence, multiple diverse measures for algorithmic fairness are known [74,75]. The term fairness was discussed in detail as early as the 1960s, with many of these original definitions being used again today [76]. As Machine Learning research has advanced, so has the importance of fairness, as the existing shortcomings have become apparent. In particular, different approaches to quantifying fairness and mitigating biases were proposed. The main distinction between those measurements is the differentiation into group fairness and individual fairness. Hereby group fairness means that persons from similar groups need to be treated similarly [77]. Hence, there should not be a bias against or towards a minority or any other group. Individual fairness on the other hand means that similar individuals need to be treated similarly [78]. Thus, people with similar attributes should receive a similar prediction from the AI system. The understanding of fairness in the technical field, therefore, differs from the legal perspective, which involves a more subjective understanding.

It should also be emphasised that these fairness concepts are fundamentally conflicting. While group fairness can in most cases be determined using statistical measure [79], whereas individual fairness needs further insights and a deeper understanding of the data at hand [75]. Hence, group fairness is largely favoured in the literature because it is often easier to implement and enforce [80]. As a combination of both approaches, Kearns et al. [81] coined the term subgroup fairness. This approach interpolates between statistical and individual fairness to achieve the benefits of both concepts. This means that biased decisions, resulting from the combination of multiple features can be identified [81]. However, these concepts represent the overarching idea of equal and thus fair treatment [82].

2.2.1. Group fairness

Group fairness was researched very early on [77] and is still a very important field of research today, which is evident from the fact that many new research approaches define fairness in terms of group fairness [83,84]. Group fairness is to be understood as a concept of achieving algorithmic fairness when dealing with multiple groups. Considering a simple case of two groups, one advantaged and one disadvantaged, specific features of these groups are considered sensitive attributes, which determine the corresponding group membership. For example, a group of people can be divided into subgroups by sensitive attributes, such as ethnicity, sex, or age. If the prediction outcomes for people belonging to one of those groups differ, a bias is introduced and one group might be privileged compared to the others. Individuals are assigned to a specific group based on such sensitive attributes. Subsequently, it is then analysed whether people in this group receive different results compared to people who do not belong to this group. If a group is treated worse as a result, there is a bias. Group fairness is based on statistical parity between certain groups (e.g. gender, age, and race). This becomes clear when, for example, women always get a different result than men or young people are favoured over older people.

However, a group bias could also be present, if not one group is unprivileged per se, but members at the intersection of two groups are treated in a different manner [74]. The study by Buolamwini and Gebru [85] highlighted this problem, where commercial gender classification systems were investigated. It was revealed that dark-coloured females were the most misclassified group, though the groups of females or people with darker skin tones in isolation did not show the same effect.

Measures for group fairness. Group fairness can be implemented through various statistical and algorithmic techniques that aim to ensure fairness at a broader level. In the literature on group fairness various approaches of measurement are proposed. Most criteria for measuring group fairness can be divided into three categories. This includes independence, which compares decision rates across groups, as well as, separation and sufficiency, which compare error rates across groups [26]. The exact focus varies between the different approaches.

Statistical parity is also referred to as demographic parity and formally defined as follows [75]:

$$|P[TP|_{G=g_p}] - P[TP|_{G=g_{up}}]| \leq \epsilon \tag{1}$$

Whereby TP stands for the true positives, and the groups G represent either the privileged g_p or the unprivileged group g_{up}. Thus, the statistical parity compares the difference in positive prediction rates up to the predefined value of the bias ϵ, to ensure similar values for all groups. The smaller the difference, the more similar both groups are treated.

A slightly modified version of this approach represents the legal notion of disparate impact, where the requirement is that the proportion of positive predictions PP is similar across groups [75]. Hence, the disparate impact compares the ratio of positive predictions between the groups. This measure can be computed as follows:

$$\frac{P[PP|_{G=g_{up}}]}{P[PP|_{G=g_p}]} \geq 1 - \epsilon \tag{2}$$

Thus, the rate of positive predictions of the unprivileged group must be similar ($\geq 1-\epsilon$) to the amount of positive predictions for the privileged group. Again, ϵ represents the bias.

Another approach that overcomes the disadvantages of the aforementioned methods is called *equalised odds*. This approach is used to analyse results based on false positive and true positive rates [86]. Mitchell et al. [74] used the term *equal accuracy* to discuss this notion of fairness. The idea behind this measurement is that predictions do not privilege a certain group if the rate of correct predictions (positives and negatives) is equal among all groups. Hence, those values can be calculated and compared for each subgroup. The difference between the true positive rate (TPR) and the false positive rate (FPR) of both groups up to the bias ϵ can be computed as follows:

$$\begin{aligned} |TP|_{G=g_p} - TP|_{G=g_{up}}| \leq \epsilon \\ |FP|_{G=g_p} - FP|_{G=g_{up}}| \leq \epsilon \end{aligned} \tag{3}$$

Whereby TP stands for the true positive rate, FP for the false positive rate of each group, p for the privileged group, and up for the unprivileged group, respectively. The smaller the differences between both groups, the fairer. Equalised odds enforce equal bias and equal accuracy in all demographics, pushing models to only perform well on the majority [87]. This becomes clearer when considering a loan approval example. The percentage of female applicants, wrongly denied a loan, even though they would have paid it back must be equal to the corresponding percentage of male applicants [88]. Vice versa, the percentage of female applicants who are given loans, even though they are not able to pay them back must be equal to the percentage of corresponding male applicants.

The simplified version of *equalised odds* is called *equality of opportunity*, where only the true positive rate is enforced [89]. This is a relaxation of the strict equalised odds, thus weaker and allows for better utility [87].

Fairness in allocation is a measure of the difference between the enrolment and allocation status of a particular use case [89]. Thus, the difference between a certain final action and the decision status of the system. For example, a historical or social bias might persist throughout the decision-making process and result in an unjust decision for a group [89]. This becomes clear when reviewing housing allocation systems for homeless people in America. People are enrolled in an allocation system, with a positive decision on receiving a spot in a shelter. However, there is often discrimination against dark-coloured

people in the sense that they are denied a home by the landlord. Thus, the recommendations from the system are not enforced, which leads to an unjust allocation of housing between the groups.

In newer works, fairness is sometimes measured using calibration, where the calibration compares whether the probability of the model matches the actual likelihood of something to occur [90]. Another approach focuses on the worst-case outcomes across groups and tries to minimise the maximum error [91].

2.2.2. Individual fairness

However, systems that are considered fair according to a certain group fairness measure can still lead to situations in which the outcome will be perceived to be unfair from the perspective of the individual. Such a scenario would be a university applicant who belongs to an ethnic minority and gets chosen over another applicant from the majority ethnicity, even though the latter obtained a better score for his admission test. While this decision might reduce ethnic disparity at the university, it would at the same time be experienced as unfair by the rejected applicant.

Moreover, AI systems can avoid detection of some forms of discrimination completely if only group fairness is considered. Suppose a bank uses an AI system to grant loans to applicants from two different ethnic groups. While the software grants loans to 30 percent of applicants from ethnicity A at random, it grants loans to the 30 percent of applicants from ethnicity B with the most savings. This software would be considered fair in terms of group fairness even though the process by which loans are granted is vastly different between the two ethnic groups [92].

To measure this disparity, individual fairness focuses on the individual rather than on the group. The aim is to ensure that similar individuals are treated similarly. In a more abstract formulation, individual fairness, therefore means that each individual is treated fairly according to their unique circumstances and characteristics [78].

Fairness through unawareness. A very intuitive approach to remove discrimination in the example of the loan-granting software would be to simply ignore the sensitive attribute *"ethnicity"* during the decision-making process and train a classifier with the remaining features. In the work of Kusner et al. [93] the term *Fairness through Unawareness* is defined as *"An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process"*.

This idea of excluding certain features from the decision-making process stems from the research of Grgic-Hlaca et al. [94] on *"process fairness"*. Process fairness is a measure which is defined as the fraction of users that deem the usage of a particular set of features as the basis for making decisions about individuals fair and appropriate. The focus of this measure lies on how decisions are made and which features are considered rather than the actual outcome.

The potential problem of Fairness Through Unawareness is that there might still exist relationships between sensitive attributes and non-sensitive attributes. Even if sensitive attributes are ignored during the decision-making process, those relationships can cause indirect discrimination [22].

A Bloomberg analysis [95] discovered that even though Amazon stated that ethnicity and demographics were not explicitly used for deciding which ZIP codes in US metropolitan areas were eligible for Amazon Prime's Free Same-Day Delivery service, most of the areas excluded from that service were predominantly non-white neighbourhoods. Cost-related data, like the proportion of Amazon Prime members in a specific area, the distance to the nearest distribution centre, the local infrastructure, or delivery partners were used to identify the most suitable ZIP codes which would result in the highest revenue for Amazon. However, this data is often heavily biased and reflects the historic racial divide within cities. Omitting the ethnicity of ZIP codes does not inherently guarantee an improvement in the fairness of predictions. Hidden relationships between features were dismissed, as sensitive features were not directly incorporated into the decision-making model. This is a prime example for *Fairness Through Unawareness*.

Unless a deliberate effort is made to purposefully identify such relationships and biases beforehand, they will inevitably influence the results of AI systems, when being trained on them. For the task of correctly identifying and mitigating such problems, domain knowledge is often required [22].

Fairness through awareness. One of the seminal works in the field of individual fairness was authored by Dwork et al. [78], in which the concept of *"Fairness Through Awareness"* was introduced. Fairness through awareness assumes that each person has unique characteristics and experiences that can influence the fairness of decisions. To achieve individual fairness, relevant attributes and characteristics of individuals have to be taken into account, including factors such as experiences, preferences or even individual histories. Ensuring individual fairness first requires a clear definition of similarity between individuals, based on transparent assumptions on the underlying processes. This is often technically realised via distance measures, where the closeness between data points is assumed to represent similarity. These distance measures are specific to the task and might be imposed by external entities like regulatory bodies. For this cause, such a measure is assumed to be public and subject to discussion and constant refinement [78].

According to the definition by Dwork et al. [78], a classifier is a mapping from individuals to probability distributions over outputs. The actual class of the individual is then chosen according to that distribution. The idea is that the output distributions of similar individuals should also be similar in order to achieve fairness.

$$D(M_x, M_y) \leq d(x, y) \tag{4}$$

M maps from the set of individuals V to a distribution over outcomes $A = (0, 1)$ for a binary classifier. M_x then denotes the outcome distribution for the individual x.

$$M : V \rightarrow \Delta(A) \tag{5}$$

D refers to an arbitrary distance measure that reflects the distance between two probability distributions. This distance must be less or equal to the distance d of the two individuals according to the distance measure [78].

The inclusion of different individual aspects poses an essential difficulty in implementing individual fairness. Furthermore, individual fairness has to be balanced with other considerations such as overall utility and efficiency [96]. However, it is important to note that this trade-off seems to exist across all fairness measures, not limited to individual fairness [97].

Considering causality. As shown in Bloomberg analysis of Amazon's same-day delivery service, biases can enter into a system rather unnoticed and create indirect discrimination [95]. It can be difficult to identify the cause of such indirect discrimination, as this may require expert knowledge in the target domain in order to identify potential (causal) relationships between variables used in the decision-making process [98].

These relationships can be made more explicit with the help of causal graphs. A causal graph is a directed acyclic graph (DAG) in which nodes represent variables (e.g. age, income) and edges the relationships between them. With causal graphs assumptions, considerations and knowledge about the data generation process can be made explicit and modelled in a more formal way [14].

Methods based on causal reasoning involve the construction of a causal graph and then checking the paths in order to detect whether a classifier trained on that data would either directly in indirectly discriminate based on sensitive attributes [26]. If a path can be found from a sensitive attribute (e.g. gender) to the target variable (e.g. hire) it is an indication of possible discrimination. This was further formalised by Nabi and Shpitser [99] as *"path specific effects"*. In their work, they measured these path-specific effects as the effect a sensitive attribute

Table 2
Comparison of Group Fairness and Individual Fairness.

	Group fairness	Individual Fairness
Protected Entities	Examines predefined groups based on protected attributes such as race, gender, age, or disability.	Considers fairness at the individual level, without relying solely on group membership.
Fairness Measures	Evaluates statistical parity, equal opportunity, fairness in allocation, and other group-based fairness measures.	Focuses on similarity-based fairness, personalised decision-making, and individual-level outcome comparisons.
Scope	Evaluates the impact of an algorithm or decision on predefined groups based on protected attributes.	Considers the unique characteristics, needs, and circumstances of individual users or decision subjects.
Considera-tions	Considers systemic biases, disparate impact, and historical inequalities affecting the protected groups.	Accounts for personal attributes beyond group affiliation, aiming to avoid discrimination and bias for each individual.
Intervention	Group interventions aim to mitigate disparities across protected groups as a whole.	Individual interventions are tailored to address the specific needs or circumstances of individual users.
Responsibility	Emphasises collective responsibility to address systemic biases and structural inequalities.	Highlights individual responsibility in ensuring fairness and avoiding bias in decisions.
Advantages	Provides a broader approach to address systemic discrimination and promote equality among protected groups.	Considers the unique needs and circumstances of individuals, ensuring personalised and context-aware decision-making.
Disadvan-tages	May overlook the variations and differences within protected groups, potentially reinforcing stereotypes or neglecting individual experiences.	Individual-focused approaches may overlook systemic inequalities and perpetuate biases if not contextualised within broader social structures.
Unintended consequences	May struggle with reverse discrimination or neglecting merit-based considerations.	May struggle with resource allocation and scalability in larger-scale decision-making systems.

exerts on the outcome over various direct and indirect paths. These effects can then be mitigated by transforming the input data.

However, these path-specific effects do not necessarily have to be discriminatory. Influences of a sensitive attribute might be deemed unproblematic they are justified. This is introduced in causal graphs via *"resolving variables"*. If a path from a sensitive attribute to the outcome is *"blocked"* by a resolving variable it is considered non-discriminatory [100].

2.3. Redefining fairness

After considering fairness it is evident, that both group fairness and individual fairness are suitable solutions for questions regarding the equality and balance of AI applications. Nevertheless, there are also some differences which are recapped in Table 2.

In addition to the technical classification, however, the legal perspective on fairness is also important, particularly with regard to the technical approaches of group and individual fairness. While from a legal point of view, fairness is primarily understood as impartiality and corresponding regulations are established to ensure it, fairness in technical contexts lies in the categorisation of individual fairness and group fairness, and thus in different approaches. The distinction found in the technical consideration of fairness fails to find its equivalent and does not align with the legal requirements. While the distinction between group and individual fairness seems to make sense from a technical point of view, it leads to inequality and injustice from a legal point of view, as it is not based on a uniform concept of a decision, but on different approaches that produce different results from the same initial situation.

From a legal point of view, however, the use of two different concepts of fairness does not appear to be appropriate. A potential strategy for addressing the tension between group fairness and individual fairness involves adopting a context-specific approach, as proposed by Binns [82]. However, this approach introduces the possibility of divergent interpretations of fairness depending on the specific circumstances, thereby potentially resulting in unequal treatment across cases. Furthermore, if fairness is perceived differently from case to case, the fundamental right of equality before the law according to Article 20 CFR and thus the consistency of the law is undermined.

Preferably a uniform and concise definition of fairness should be used in all application cases in order to create comparability. However, this might not be possible due to the multifaceted understanding and divergent use cases. For example, Google's algorithm accused a father of child abuse due to images of his child's private parts, uploaded to the healthcare provider's messaging service [101]. The images were automatically uploaded to his Google Cloud and analysed according to Google's terms. Hereby, the father was not only accused of a heinous crime, he did not commit, but was also denied any further service of Google since this was against their policy. This shows the immense risks that AI might pose to individuals, and that fairness as such cannot be ensured by rules. The case of an algorithm mistakenly identifying an individual as a child abuser serves as a stark example of such limitations. Of particular relevance in this context is the ongoing debate on how to legislate on these issues and define fairness, which highlights the need for a comprehensive approach to understanding and mitigating fairness-related risks in AI systems.

The question that emerges is which fairness concept, from the perspective of legal sciences, should be favoured and which one fulfils the legal requirements. Applicable law requires that every individual person is to be treated equally. Moreover, the focus is not on a specific group but on the individual herself. According to this, fairness would also have to be seen with regard to the individual. With reference to Article 20 CFR, it can be argued that individual fairness corresponds to the intention of the legislature.

Another approach to interpret fairness can be derived from the categorisation of the European court system as well as the court system in most European countries as civil law. Civil law systems trace

their origins to the legal principles and codes established by the ancient Romans, which follow clear procedural rules and use formal documents [102]. These systems are characterised by comprehensive legislative codes that serve as the primary source of law. In civil law systems, legal decisions are typically based on the interpretation and application of these codes, rather than relying heavily on prior court decisions as in common law systems. This leads to the fact that each court bases its decision on the same procedural principles and thus each case is considered individually, but the decision-making is uniform. This ensures that, on the one hand, the rights and obligations of the parties before the court are clearly defined and, on the other hand, that no disadvantage arises for individual parties. According to Burdick [103], these principles of Roman law, which ensure fairness and equity, are the reason why many legal systems are based on them. The Roman praetors, who were responsible for the jurisprudence, often used the phrase *"ex bono et aequo"*, i.e. *"right and fair"* [104], and the Digest of Justinian also define law as the *"art of what is right and fair"* (Iustinianus, Digesta Iustiniani 1.1.1.pr.1). Civil law is often perceived as fairer due to its structural characteristics and the inherent ability to reevaluate court decisions. In this legal system, influence over the law extends beyond the legislative branch and includes the judiciary. As a result, subsequent cases are assessed based on prior legal rulings. However, it is important to note that revisions of such precedents are possible in all cases, meaning that unlike in common law systems, judicial precedent is not inherently binding.

In this sense, it is not the decision itself that is the decisive element of the civil law system, but the uniformity of the processes. It also follows that the structure of the system contributes significantly to fairness. In his decision, the judge has discretionary powers to interpret the facts of the case, but the solution of the legal question and the associated legal assessment of the facts are standardised so that each party is entitled to his rights and fairness is guaranteed.

Linking back to the aspect of technical fairness, the aim hereby is to design the decision-making process to be uniform and consistent, similar to the civil law system. The goal is to ensure that all individuals or entities subject to the system are treated fairly and consistently, regardless of their individual characteristics. Just as the civil law system emphasises uniformity of processes, technical fairness aims to provide consistent and standardised treatment to avoid bias and discrimination. This standardisation ensures, that that each party is entitled to their rights and that fairness is guaranteed.

3. Bias mitigation

Data is the core component of a predictive model. The data used for the training of a model might introduce undesirable properties when being used as a basis for decision-making [74]. Such properties are generally labelled as bias [105], although it can be distinguished between statistical and societal bias. According to Mitchell et al. [74], statistical bias is a more precise term, which describes non-representative sampling and a measurement error, whereas societal bias represents social structures and past injustices within the data. If biased data is used to train a predictive model, such undesirable properties might be exacerbated. Thus, it is of utmost importance to detect a bias and mitigate its effect on the model.

3.1. Types of bias

Different types of biases can occur due to the origin of the data, the chosen processing steps and methods, or even the selection of training and evaluation metrics [106]. For example, feature hunting [107] is a greedy approach to testing multiple features for classification tasks until finding a feature with the highest improvement rather than testing features based on a hypothesis. Hellström et al. [108] proposed a taxonomy of different types of bias that may appear in the process of generating a machine learning model. This taxonomy does not only include bias within the data but also focuses on bias introduced due to historical or social norms, learning bias of the model and evaluation bias. The essential part is that the authors proposed new terminologies such as a *specification bias* for the choices and specifications of inputs and outputs of a training task. This illustrates that different levels of granularity are understood within the scientific community. Regarding different sciences, other biases can be observed. For example, behavioural bias, or even position bias, leading to different interactions and intentions of the user could be discussed. However, this article focuses on those types connected to machine learning and fairness, hence the following definitions.

There are many different kinds of biases. Commonly mentioned biases in the literature include the following:

1. **Algorithmic bias**. Algorithmic bias exists when there is no bias present in the data, rather the bias is added solely by the algorithm and its corresponding design choices [109]. As a result, models may fail to treat groups fairly under given conditions [110]. Moreover, AI systems lack the human inherent *common sense* that guides human decision-making, exacerbating these biases. Their inability to fully understand or appropriately prioritise contextual information can further skew outcomes [111], and current AI fails to distinguish between correlation and causation and requires appropriate training data [112]. The key problem hereby is that while computational methods can recognise and potentially mitigate bias in data sets, they often fail to capture the socio-cultural or ethical complexities underlying bias and discrimination.

2. **Historical Bias**. Historical bias is the already existing bias in the real world and socio-technical issues [109]. Even if the data is perfectly measured and sampled, the usage of this data can lead to a model that produces harmful outcomes [113]. Historical biases are often linked to gender. For example, fewer women were in leading positions, resulting in fewer search results for 'female CEOs' than male ones [106]. However, not only gender is linked to this kind of bias, but also ethnicity, cultural or social constructs [114]. Those biases might represent the reality, however, it should at least be considered, if such a bias should be reflected by an AI system.

3. **Representation bias**. Representation bias arises from the manner of data sampling from a population. This results in non-representative samples or unbalanced data with a lack of diversity, missing subgroups or other anomalies [109]. Such deficiencies may result in unjust predictions towards minorities. For example, datasets like ImageNet exhibit a lack of geographical diversity, which results in verifiable bias in favour of Western cultures [109].

4. **Sampling bias.** Sampling bias is similar to representation bias [109]. This bias occurs when the training data is not representative of the target population. Thus, certain groups are either under or over-represented in the training data, compared to their actual prevalence in the real-world population. This bias can lead to preferential treatment of over-represented groups or discrimination against underrepresented groups. Further, the trend estimated from one population might not generalise to data from a new population [109].

5. **Measurement Bias**. Measurement bias arises from how features are chosen, utilised and measured [106]. Features or labels are typically seen as proxies, which are chosen to approximate a specific construct [113]. If those proxies are poor reflections or the target constructs are computed differently across groups, these proxies become problematic [113]. Hence, this bias results

from using mismeasured proxy features [109]. For example, COMPAS [9], the recidivism risk prediction tool predicts the risk of an individual based on features such as prior arrests and friend/family arrests [109]. Minorities often represent groups with higher arrest rates, due to the fact that they are policed and controlled more frequently. Hence, the method of measurement across groups varies, resulting in a measurement bias that links individuals of minorities with higher arrest rates.

6. **Omitted Variable Bias**. Omitted Variable Bias occurs when important features are not considered by the model [109]. For instance, a model that predicts the credit score of an individual needs to incorporate all relevant information in order to return accurate predictions. Therefore, the model needs to consider not only the salary and debts of a person but also the heritage or other assets, to assess the full extent of creditworthiness.

7. **Aggregation Bias.** Aggregation bias occurs when assumptions about individuals are drawn from the entire population [109]. The underlying assumption hereby is that the mapping from inputs to labels is consistent across groups [113]. This includes the generalisation of assumptions between subgroups. Thus, if there are significant differences between groups they should not be generalised and used for each individual. For example, there is a difference in the compatibility of medicines between children and adults. This is an important feature, which should not lead to a general assumption, such as equal tolerance for every patient regardless of age. Aggregation bias can result from any general assumption about a subgroup within a population [109], and it can lead to a model that is not optimal for any group [113].

8. **Evaluation Bias**. Evaluation bias occurs when a benchmark dataset that does not represent the target population is used for the evaluation of a model [110,113]. Models are optimised on their training data, but the quality of a model is often assessed with the help of benchmark datasets [113]. Benchmark datasets allow for an objective comparison between machine learning models. However, if the benchmark dataset is not representative of the target population, models which only work well on this subset of data could be preferred [110]. For example, Buolamwini and Gebru [85] showed that the underperformance of dark-coloured females was overlooked due to an underrepresentation in benchmarking datasets.

9. **Popularity Bias**. Popularity bias occurs in recommender systems when frequently rated items get more exposure, and are thus recommended more often than less frequently rated items [115]. Depending on the user's preferences these recommendations might be not appropriate regarding their real interests [115]. This bias might seem unproblematic, however, e-Recruiting recommender systems might be used to recommend applicants to a recruiter, based on their profiles [116]. Whereby the profile or specific attributes of applicants might amplify this bias and lead to an unfair distribution of exposure.

While this list of biases may not be comprehensive, it does demonstrate the diverse range of biases that exist. Regarding *unfair biases* mentioned in the AI Act, it is important to revisit this term. Naturally, this brings up the question of which biases should be considered unfair. In principle, it seems reasonable to assume that any bias is negative and can thus be considered unfair, which is not entirely true. A bias is not negative in every case, but there are also cases in which a positive bias is spoken of [117]. An example of this is the preferential selection of women in the application process to increase the quota of women in certain occupational areas. Having a preference can indicate bias, but not all biases are necessarily unfair, as it depends on the underlying intention. The legislator's statement that unfair biases should be prevented is difficult to understand from a technical standpoint without first providing a clear definition of fairness. Exemplary historical and social biases exist everywhere and often go unnoticed, yet they lead to unequal treatment [118]. However, it is impossible to completely eradicate these fundamental inequalities of society as they are inherent to human nature [119]. In practice, therefore, the question arises as to which criteria should be used to determine when a bias is unfair. This further underlines the importance of a clear definition of fairness, as it is only on the basis of this definition that it is possible to determine the existence of an unfair bias.

3.2. Bias amplification

Most automated decision-making or recommender systems have been introduced with the promise of reducing human error while providing more accurate results [120]. However, it is precisely this assurance that changed the way people approach their work and introduced new errors. Automated decision-making systems (ADMs) gather and process data in order to make qualitative decisions with minimal to no human intervention [121]. This implies that crucial decisions that impact one's livelihood, environment, or even life are taken by the AI system instead of a human. ADMS are responsible for making evermore critical decisions and are thus used in areas such as autonomous driving cars [122]. For example, in the case of an accident, a self-driving car makes a life-or-death decision either in favour of the driver or other involved parties.

Recommender, or automated decision-support, systems try to predict future user interests, based on previous interactions or gathered data about user preferences [123]. Famous examples of this are the recommended products on Amazon or the movie recommendations on Netflix [123].

Further, it should be noted that intended decision-support systems might become decision-making systems when used incorrectly. For example, Apple watches provide the functionality to warn users of an irregular heart rhythm. However, most users rely on these apps to the point that they do not consider signals from their own bodies if the app fails to alarm them [124]. Thus, they may trust certain decision-support systems to the extent of ignoring their instincts, promoting these systems to decision-making, rather than decision-support systems.

A particularly striking case for an automated decision-making system is the use of weapons, where for example the Israeli military utilises an AI system called *"Lavender"* for targeting. This system scores individuals in Gaza based on their perceived likelihood of being Hamas militants, utilising extensive surveillance data. Reports indicate that human oversight of these systems is minimal, often reduced to a mere formality, with decisions made in seconds, focusing on male targets for potential action [125].

The study by Skitka et al. [126] showed that people might fail to respond to irregularities or events because automated devices failed to recognise or indicate them. In addition, their experiment showed that participants are more likely to follow the automated directive, although provided with contrary information from a reliable source. This notion is called *automation bias*, which describes the tendency of using automation as a heuristic replacement for vigilant information seeking and processing [127]. Studies showed that participants without automation aid outperformed those with an automation aid [127]. Furthermore, participants are more likely to act on automated recommendations, regardless of their validity. Hence, automation bias in decision-support systems might result in an instant acceptance of the decision instead of critically questioning.

Feedback loop. An additional problem of unresolved bias within an AI system includes the pernicious feedback loop it creates [128]. If a system disadvantages a group of people based on an assumption such as, *"good employees are those with a higher credit score since they pay their bills on time and follow the rules"*. This also implies that people with low credit scores are less reliable and thus might have more trouble finding a job. That exact reasoning makes it harder for a person with a low credit score to find a job, which would help them pay their bills, and even worsen their current score. These feedback loops help the environment to justify their assumptions without being concerned about the fates of individuals [128]. As a consequence, this feedback loop continues and thereby makes the model more unfair with every decision. Another example mentioned by O'neil [128] about feedback loops is represented by mortgage securities, which were before considered simple financial instruments. However, these developed into frauds, where customers were able to lend money for homes they could not afford. Unsustainable mortgages were written, the fees were collected and the resulting securities were sold. This, of course, was very profitable for the banks, who were ruining thousands of people. The perfidious thing was that they especially targeted poor and minority neighbourhoods. Thus, getting rich at the expense of others. The risk rating of those securities was opaque, thus leaving people to believe they would get their money back, including interest, when buying these securities. Banks held onto false underlying assumptions, namely, that not many people would default at the same time, and that the risks are carefully balanced. Investors bought mortgages with underestimated risk, leading to costly consequences. This is an example of a model, that favoured customer satisfaction over the accuracy of the model, incorporating a pernicious feedback loop.

3.3. Bias mitigation techniques

Bias mitigation techniques can be applied in various stages of the machine learning pipeline, e.g., data collection, pre-processing, in-processing, post-processing, or even after evaluating a machine-made decision [113]. A phase of careful planning before building a dataset or a model can help identify and mitigate risks [129].

Data collection or data generation is the initial step in every machine-learning model. Hence, it is crucial to be aware of the data and the bias it may hold. Identifying possible hazards within the data requires knowledge about the domain and the intended use of the system [113]. Thus, possible tendencies resulting from a historical bias, representation bias, sampling bias or measurement bias can be prevented before they become a problem. Nevertheless, not only the data might hold a bias, but the design choices of the model itself can contribute to algorithmic bias [130]. Realising that algorithms are not impartial and how the design of a model can affect the outcome is crucial to mitigation techniques [130].

The list of methods presented here is not comprehensive and should not be interpreted as a ranking of their significance. Instead, it aims to give readers a general idea of different methods and strategies used. The methods included in the following were chosen based on their widespread use.

3.3.1. Pre-processing

Pre-processing addresses a main cause of bias, namely an unbalanced dataset before the model is actually trained with the data at hand. Those techniques assign different weights to the training data, based on the categories of sensitive attributes or outcomes, remove sensitive attributes or changing labels to remove the bias [131,132]. Simply removing sensitive attributes might not work well since group differences can persist due to proxy values. Meaning that group membership can be reconstructed from different features, which in turn allows the algorithm to categorise the data points into groups [13].

Massaging and reweighting. Calders et al. [77] proposed two approaches to remove bias from a dataset. Hereby, the authors concentrated on unjust dependencies between data attributes and class labels. The authors state that such problems could arise, if the training data is gathered from different sources with different labelling criteria, or when the data is generated by a biased decision process. The first solution to remove such a bias is called massaging. Massaging removes the dependency between the attribute B and the attribute class by changing the labels of some objects $x(B)$ equal to the value b from '−' to '+'. For the same amount of objects $x(B) \neq b$, the label will be changed from '+' to '−'. This approach changes the labels of the objects and is thus rather intrusive. The second approach proposed by Calders et al. [77] attaches different weights to the objects instead of relabelling them. For example, objects with $B = b$ and $class = +$ will get higher weights than similar objects with the class '−'. On the other hand, objects with class '−' and $B \neq b$ will get higher weights than similar objects with class '+'. The objects are then sampled according to these weights, resulting in a dataset without dependencies.

Data augmentation. One pre-processing technique that can be used to mitigate the effects of bias within a dataset is called data augmentation [133]. Hereby a new dataset is created, swapping the protected attribute to even out the differences, while eliminating the correlations attributed to the groups. For example, consider a hiring algorithm and the underlying dataset. If there are more males than females present in the training set, the dataset is unbalanced and might result in a disadvantage of females, while promoting males. In order to counteract this behaviour, the training data is duplicated, and all male entities are swapped with female ones and vice-versa. This results in the aforementioned elimination of gender-based correlations, while maintaining non-gender-based correlations. Thus, the gender, or the initial lack of uniformity between both groups is no longer able to directly influence the predictions, made by a system.

3.3.2. In-processing

In-processing techniques are used during the training phase of a model [131]. Whereby, the fairness constraints are viewed as a constraint optimisation problem, hence they are transformed into penalties [134]. Nevertheless, this approach might lead to an unstable training process or introduce unnecessary complexity [135]. Oversampling or adversarial debiasing is an in-processing method, which forces the model to account for underrepresented groups to achieve a better overall performance [132].

Adversarial debiasing. In addition to introducing bias through the training data of an AI system, it is also possible to introduce post-training bias during the application of the system. For example, adversarial attacks can be used to significantly influence the results of such systems. Hussain et al. [136] showed that adversarial attacks aimed at degrading fairness can significantly degrade the fairness of predictions with a low perturbation rate and without a significant drop in accuracy. At the same time, this method can also be used for bias mitigation. Adversarial debiasing is a technique, whereby classifiers and their corresponding adversaries are trained simultaneously [137]. Both models perform against each other to increase the performance, while decorrelating the sensitive attributes from potential biases [138]. While the classification model tries to predict the ground truth, the adversarial model tries to exploit fairness issues. This helps the learner to identify possible problem areas, which are optimised. Further, adversarial models can be used to not only increase the group but also the individual fairness [139]. This approach provides the benefit of generalise-ability across different datasets and applications, while not requiring any prior assumptions about the distribution [138].

Compositional. Compositional approaches tackle bias by training multiple classification models in order to utilise specific models to reach predictions for different population groups [137]. These models can also be used in an ensemble fashion. Decoupled classification models for different protected groups can achieve an increase in accuracy, however, this also entails a reduced amount of training data for each classifier. Various concepts try to counteract this disadvantage by introducing transfer learning and different weights for each classification task [140]. As a result *preference guarantees* [141] determine that a decoupled classifier is preferred over a classifier trained on all data, or any other classifier by a certain group. A slightly different attempt, *fair use*, was brought forth by Suriyakumar et al. [142], which specifies that a performance improvement for every group should be attained if a classifier uses sensitive group information. This advance resulted in the training of multiple classification models with different fairness goals. For example Liu and Vincente [143] conceptualise bias mitigation as a multi-objective optimisation problem, which studies fairness-accuracy trade-offs under consideration of multiple fairness metrics. Hence, practitioners are able to choose an approach, depending on the fairness-accuracy trade-off, that suits their use case at hand.

Autoencoders. An autoencoder is a neural network, which is trained to reconstruct its input [144]. Thereby a representation of the data is learned through an unsupervised manner. This means the input is encoded into a compressed and meaningful representation, and afterwards reconstructed such that the output is as similar to the input as possible. However, these models can also be used to eliminate bias during the training process. Variational autoencoders are used in recommender systems, to counteract popularity bias [145]. These generative models are based on Multi-Layer Perceptrons (MLP), and recently gained popularity as a strong method to implement collaborative filtering recommendations [146]. Variational autoencoder infers latent representations of interactions as normal distributions, where the parameters are estimated during the training process [145]. In addition, Grari et al. [147] proposed to use causal graphs from the underlying data to create latent representations, containing information about the sensitive attribute. Afterwards, a model is trained using these proxies in combination with penalised adversarial during the training process. Hence, a predictive model is trained without the availability of sensitive demographic information during the training phase. Tyagi et al. [148] state that an enhanced variational autoencoder can be used to mitigate gender bias while preserving gender-related information. They showed that an enhanced variational autoencoder can mitigate the bias of pre-trained static word embeddings while preserving syntactic, semantic and gender information. Hence, this approach achieves non-biased representations of gender words.

3.3.3. Post-processing

Post-processing steps to enforce fairness are decoupled from the training process [131]. Generally, model predictions are manipulated on the basis of fairness constraints, which makes it theoretically possible to apply this method to any machine learning model [138]. In addition, there is no need to access model parameters or training data with this approach. However, fairness is only increased with respect to the specified constraints, not necessarily to any other fairness notion. Although this process is likely to decrease the bias of a classifier, the accuracy fairness trade-off might not be ideal [88]. Moreover, post-processing poses limited flexibility, insofar that notions such as equalised odds are not achievable with deterministic solutions, due to the fact that those allow only for single error constraints [90].

Equalised odds. Equalised Odds is a notion that allows checking for discrimination with respect to specified protected attributes [87]. It enforces equal true and false positive rates in all demographics. Hence, models that perform well only on the majority are punished. This technique discovers probabilities, with which to change output labels such that the requirement of equalised odds can be satisfied. Only the aggregate information is needed to execute equalised odds as a post-processing step. Thus, this could also be done in a privacy-preserving manner. [90] proposed calibrated equalised odds to reach the equality of false negative and false positive rates between the groups. Therefore, a calibrated probability estimate is taken into consideration.

Reject option-based classification. Reject option-based classification (ROC) exploits predictions with high uncertainty, and assigns predictions with favourable outcomes to unprivileged groups and predictions with unfavourable outcomes to privileged groups [149]. The underlying assumption hereby is that the most discrimination occurs when the model is least confident about a prediction [4]. Hence, predictions of individuals close to the decision boundary are modified. However, this might lead to the introduction of new biases, which favour unprivileged groups.

4. Open challenges and future research directions

4.1. ChatGPT: A contemporary example

One of the best-known AI tools at the moment is ChatGPT, which at times triggered a mass hype around AI tools due to its groundbreaking ability to respond to user input. This hype also led to the advance of the scientific discourse on the use of AI in society [150–152]. Basically, ChatGPT is a Large Language Model (LLM), which is pre-trained by huge amounts of data, such as webpages, books or other written material and deployed as a conversational AI system. Generative Pre-trained Transformers (GPT) models are designed to generate natural language text that is coherent and consistent with human language [114]. The pre-training allows the model to learn patterns and relationships between words in natural language. Combined with the sheer amount of data gives LLMs the ability to effectively generate coherent and realistic answers. Further, this model is able to perform a wide range of natural language processing (NLP) tasks, such as text generation, question answering, language translation, and sentiment analysis.

It should be emphasised that ChatGPT is not the only language model that has been developed. There are already numerous other AI systems like BARD and Watson [153], as well as open-source model such as Llama 2 [154]. Nevertheless, the advent of ChatGPT has raised several ethical and legal questions. The origin and content of the training data used, including the possibility of copyright infringement [155]. The same applies to the problems of medico-legal issues [156] and plagiarism [157]. Other aspects relate to issues of data protection, as illustrated by the temporary blocking of ChatGPT in Italy [158]. These problems, however, are by far too extensive to be covered in a single paper.

Furthermore, ChatGPT showed remarkable abilities to generate human-like text, although the information may sometimes be misleading or simply incorrect [159]. A major challenge is to ensure and maintain the accuracy and reliability of the AI-generated context. This also entails a requirement of quality control for the system. In addition, these models are by no means immune to bias. For example, possible biases within the training data could be picked up and amplified by the system [160]. If used in critical areas such as healthcare, employment or within the legal system, these biases could have disastrous consequences. The examination by Touvron et al. [154] highlighted that different models demonstrate disparate sentiments towards sensitive attributes. Hence, ChatGPT overall tends to have a more neutral sentiment score, whereas a fine-tuned version of LLAMA 2-Chat tends to show more positive sentiment. The results of the study indicate that LLMs exhibit a more favourable attitude towards American female actresses than towards American male actors. In addition, a study by Rozado [161] has shown that ChatGPT responses are biased towards a political agenda. Thus, without clear transparency policies and measures to mitigate bias as much as possible, major problems will remain and have the potential to manipulate human decisions.

4.2. Takeaways for lawyers and policymakers

Although research has already proposed different approaches to ensure fairness, no legal guide is in sight so far. Rather, the legislature chooses general and unclear formulations, which appear to be appropriate and solution-oriented on the surface, but on closer inspection often raise questions without offering any answers. In this sense, reference should be made to the *General Data Protection Regulation* (GDPR) [162], which, despite its introduction in 2016, still leaves many things unclear and presents developers with almost insurmountable tasks [163]. The GDPR also incorporates principles of fairness, especially in Article 5(1)(a), where it clarifies that personal data shall be *"processed lawfully, fairly and in a transparent manner"*. This mostly addresses the aspect of procedural fairness, however, substantive fairness is not taken into account. The focus on procedural fairness inadequately addresses the nuances of data processing's substantive fairness, leaving gaps in effective data subject protection. Substantive fairness, encompassing principles of different areas of law such as good faith, respect for autonomy or data accuracy, are not explicitly defined within the GDPR. Many arguments can be found in the literature, as to why substantive fairness should also be an essential component of legislation [164].

Given the shortcomings of the GDPR and in view of the extensive possibilities of using AI systems and the associated danger for people, it is not advisable to regulate AI in this way. With regard to the problem of fairness, the importance of a clear legal regulation is particularly evident, which allows the developers of such systems to make them legally compliant and gives users the certainty that the use of such systems is clearly regulated and the fairness of these systems is ensured. For this, however, it is not enough to decide on a single definition of fairness, but rather it is necessary to deal in detail with the technical possibilities and the associated shortcomings. It is obvious that neither individual nor group fairness can actually be fully guaranteed in an AI system with the current technical methods. Rather, these methods are only state-of-the-art ways of reducing unfairness. Regardless, the aim must be to further strengthen this crucial research area in particular and to give it significant importance in order to compare the development of AI systems with a corresponding development to ensure fairness. Consequently, other approaches beyond the technical possibilities must also be pursued in order to offer people the security that fundamental rights give rise to. The possibilities of audits should be considered and security measures should be implemented in every development step during the development of AI systems, as suggested in the *Ethics Guidelines for Trustworthy Artificial Intelligence* [165], for example.

In this sense, it should also be mentioned that fairness is only a first starting point since problems such as proxy discrimination [166] are difficult to prove with the current technical possibilities and combating this with the existing legal tools is not promising. In view of the existing problems, the discussion in the European Union about the proposed *AI Act* [66] can only be seen as a first step towards solving the problem and is far from being suitable for countering the risks and dangers posed by such systems.

It should also be kept in mind that there are situations in which group fairness leads to better results, in comparison to individual fairness. This is evident in socio-political issues such as gender equality. The under-representation of women in STEM subjects is particularly worth mentioning [167]. The evaluation of such socio-political inequalities can only be evaluated from the point of view of group fairness, as this allows for an overview of the population as a whole, rather than focusing on the individual. These issues of social policy, however, must be separated from legal fairness and, consequently, from fairness in AI systems. The aim here is not to ensure equal treatment of the individual but to ensure the social organisation of society. A strict separation is therefore essential in this matter. Society's injustices are not the responsibility of the judiciary. Rather, it is the task of politics to enforce social change through other means such as incentives. Only then can social equality and thus social peace be achieved.

4.3. Takeaways for technical implementation

The implementation of the legal requirements for AI systems is very challenging, as in addition to the technical implementation, extensive legal requirements must be fulfilled. These require an in-depth analysis of existing legal provisions as well as an understanding of legal implications and issues. The example of fairness clearly illustrates this challenge in implementing such systems. Fairness is understood differently in the legal sense than in the technical context and thus this difference in definition leads to significant problems in the development of AI systems [168]. If the legislator stipulates that AI systems must be fair, without specifying a concrete set of criteria, from a technical point of view this leads to the fact that a variety of fairness methods can be chosen, which, nonetheless, produce different results [169].

Ensuring the fairness of an AI system should also be seen as a key consideration of developers and ML engineers. For this, it is also necessary to deal with the legal framework in detail and, as the example of fairness clearly shows, to have a certain understanding of the law. It is by no means enough to focus exclusively on the technical conditions and ignore the legal requirements. Rather, it is indispensable to take the legal components into account in every development and implementation phase of AI systems. Likewise, the Ethics Guidelines for Trustworthy Artificial Intelligence [165] recommends continuous evaluation throughout the life cycle of the AI system to guarantee safety as well as legal conformity. An evaluation of this kind, however, involves a great deal of effort and requires specialists from different fields in order to obtain in-depth results and thus discover any problems and shortcomings at an early stage. This already begins in the phase of designing the AI system, where the selection and design of the algorithm and the features used, as well as the selection of the training data. In addition, the datasets used for training should be carefully documented at the stage of their collection [129]. Every aspect of development must be carefully considered and checked against the legal requirements and correspondingly appropriate audits are to be implemented [170].

The problem here is that it is by no means easy to put legal requirements into practice. Particularly in the area of AI, the legislator limits itself to establishing basic regulations without defining them concretely. Although this allows a certain amount of freedom for development, it also leads to legal uncertainty, which affects small companies and start-ups in particular, which cannot afford lengthy and costly legal disputes [171]. As the GDPR limits certain types of data-centric innovations due to its strict data protection requirements [172], it also stimulates innovation in areas like privacy-enhancing technologies and secure data management solutions. Similarly, upcoming AI regulations may inhibit innovation by prohibiting certain use cases due to their inherent risks. However, it may also incentivise innovation on Trustworthy AI, leading to more research on fairness and discrimination. This is evident with regard to evaluation, where assessments are considered useful and necessary, but there are no standardised procedures. This lack of security must be addressed during development through clear documentation of all development steps, continuous monitoring and evaluation of the AI system, and concrete security measures and data handling practices [173]. To this end, it is necessary for development teams to be interdisciplinary in order to be able to take different aspects into account and to be able to recognise the various problems on the one hand and to find suitable solutions on the other. A key factor here is that experts also acquire knowledge of topics outside their field, so that technicians, for example, also deal with legal issues and lawyers acquire basic knowledge of data processing and data science methods.

4.4. Takeaways for science

Though there is a great deal of knowledge about fairness from both a legal and a technical perspective, the interdisciplinary debate in particular has received less attention than needed [174]. In particular, there is a lack of detailed discussions on the concrete design of fairness and its implementation in technical terms. Differing interpretations in the various disciplines are particularly problematic, as they leave room for interpretation, which in turn leads to different outcomes. Battling this uncertainty should be one of the most important research issues in the coming years, in order to guide the ongoing development of AI systems in a way that is consistent with the needs of society. The innumerable problems of bias clearly show an immense need for research to address this issue [72]. It is essential to first research on an unified definition of fairness. However, reaching a unified definition is a challenge within the legal context due to the multifaceted nature of laws and the diverse principles they encompass. Different areas of law, such as non-discrimination, rule of law, data protection, consumer law, and competition law, approach fairness from unique perspectives, reflecting their distinct objectives and societal values. Therefore such a definition would need to encapsulate principles that are universally applicable across these diverse legal areas while being flexible enough to accommodate their specific contexts. Despite these challenges, focusing on non-discrimination law as a basis for defining fairness in AI can be a strategic starting point to unify the development and use of AI systems. Only on this basis can further research proceed efficiently, since this fairness definition can then be implemented in practice as a technical solution. The current fragmentation of research approaches makes it possible to develop different solutions for different scenarios, but these do not appear to be compatible with the legal requirements. In fact, this fragmentation leads to the further dissolution of fairness and prevents the comparability of systems [175]. Although fairness is not an easy concept to define, it is necessary to find an agreed definition on which to base further analyses and comparisons. This is the only way to ensure and verify that AI systems comply with legal requirements.

Furthermore, it is necessary to find solutions to the new forms of discrimination, such as proxy discrimination, by AI systems [176]. Thus, research on bias mitigation approaches is a significant field in the near future to ensure that inequalities neither enter nor are reinforced by the AI system. The countless examples of such implicit biases clearly show that there is still a considerable need for research in this area. Additionally, the increasing complexity of the systems and the huge amounts of data have significantly increased the risk of such biases [177]. While it is no longer possible for people without basic technical knowledge to understand the results of AI systems and how they work, the increasing performance of these systems is also increasingly pushing domain experts to their limits. The sheer volume of data makes manual review increasingly impossible, and as a result, more and more biases go unnoticed. This is compounded by new problems such as proxy bias. In order to counteract this development and to guarantee transparency and traceability, new innovative approaches must be found, which on the one hand can detect and eliminate existing biases at an early stage and on the other hand can counteract the emergence of new biases, such as the automation bias. AI research must therefore place a significant focus on the issue of fairness and bias in order to confront emerging problems with solutions. Only in this way can a safe handling of AI take hold in society.

Further relevant problems that require critical engagement by the research community to achieve Trustworthy AI include the following:

1. **Data Protection and Privacy Preservation**
   The increasing performance of AI systems leads to the fact that data is processed faster and the volume of processed data is growing strongly. As a result, the privacy of individuals is becoming more and more difficult to maintain, as it is becoming easier to establish links between the data and a specific person due to the immense amount of data. Seemingly impersonal data, in combination with other information, is capable of revealing a great deal of personal information about a person and therefore poses a great threat to data protection. For example, the use of certain words or sentence construction can be used to determine the author [160]. This raises the question of how the privacy of individuals can be efficiently ensured in the age of algorithms and which methods offer adequate protection for the individual. Recent research has focused on various aspects of differential privacy, including evaluating models, balancing privacy and utility, applying methods to new threats, generating synthetic data, and using deep learning models in various applications [178].
   In addition, there are further problems arising from a data protection point of view. For example, the GDPR does not currently categorise emotional data as a special category of data, which leads to a protection gap due to its sensitive nature. The GDPR's inability to adapt to rapid technological advances makes compliance difficult for affective computing developers due to vague guidelines, particularly in relation to the processing of biometric and physiological signals for emotion recognition [179]. Furthermore, the growing volume of data and increasingly powerful AI systems pose risks to data subjects as it becomes easier to establish connections. This poses risks not only to the fundamental right to data protection itself but also to related data subject rights, such as the right to rectification or erasure under the GDPR. This creates a difficult environment for aligning developments in affective computing with EU legal standards and emphasises the need for clearer regulation and better protection for individuals' rights in an era of advanced data processing capabilities.

2. **Transparency and Explainable AI**
   Transparency, along with fairness, is seen as a key aspect of people's acceptance of AI systems. However, transparency does not only require information on how a certain system works and what purpose it serves but rather people must be given a clear understanding of how a decision is made, which parameters are used for this and how this decision leads to further consequences. Despite existing methods, such as SHAP [180], LIME [181] or DiCE [182], it is currently not possible to fulfil people's need for transparency as without prior technical knowledge, it is not possible to understand the functionality and interpretation of the information provided by the explanation methods.

3. **Interaction between Humans and Agent**
   In the realm of artificial intelligence research, the interaction between humans and agents stands as a complex and multifaceted challenge [183]. As technology continues to evolve, the integration of agents into various aspects of our lives has become increasingly prevalent, from industrial settings to domestic environments. However, achieving seamless and effective interaction between humans and agents remains a significant hurdle. Moreover, trust and cooperation are critical factors in any human–machine interaction. Building trust between humans and agents requires consistent reliability, transparent decision-making processes, and mechanisms for handling unexpected situations [165]. Without this trust, individuals may hesitate to rely on or collaborate with robots, limiting the potential benefits of their presence.

4. **Policies and regulations for AI**
   The legal debate on the functioning and effects of AI systems is only just beginning [184]. Even if the first attempts at comprehensive regulation are being made by the legislature, it is clear from the continuing progress in AI research that many

topics have not yet been discussed in deep and that there is a need for further regulation. Regulating those areas that pose particular risks to people is especially necessary. In particular, the area of fundamental rights should be emphasised, including the right to data protection and the right not to be discriminated against. The potential dangers and risks associated with AI systems necessitate a comprehensive technical examination and in-depth understanding of underlying legal problems to equip policymakers with adequate guidelines for regulating AI systems.

5. **Rethinking the concept of discrimination**
The disruptive potential of AI systems poses a significant threat to society [185]. In particular, these are new forms of discrimination which have not yet been recognised as such by the legislator or the courts. In addition to the problem of proxy discrimination, historical and social bias are also significant dangers that are difficult to overcome with the current technical possibilities. In this context, aleatoric uncertainties and unknown unknowns are particularly noteworthy. Aleatoric uncertainties stem from inherent randomness or variability in data [186]. Natural language processing (NLP) is a field of artificial intelligence concerned with the interaction between computers and human languages. One of the major challenges in NLP is the ambiguity of language, which leads to the introduction of aleatoric uncertainties. These uncertainties may result in biased predictions or decisions. Consequently, it is crucial to develop computational models that can effectively handle these uncertainties and provide reliable predictions and decisions. Failure to do so may have significant implications in different domains, including business, law, and healthcare. In contrast, unknown unknowns are risks that are completely unforeseen and unpredictable. In complex AI systems, such unknown unknowns can potentially cause inherent biases and lead to discriminatory behaviour [187]. It has been observed that machine learning models can display biased behaviour when presented with novel scenarios that were not part of their training data. This can occur due to a range of factors, including incomplete or inadequate training data, or the presence of underlying biases in the data that the model was trained on [188]. As such, it is of the utmost importance to remain vigilant and take proactive steps to identify and address these potential sources of bias in order to ensure the accuracy and fairness of the models.

5. Conclusion

This paper demonstrates the complexity and multidimensionality of ensuring fairness in AI. The main challenge of the current research is the search for a uniform definition of fairness, which on the one hand underlies the legal requirements, especially with regard to non-discrimination law, and on the other hand, corresponds to the technical possibilities. Derived from the legal requirements for fairness, which result from international law, EU law and the basic principles of case law, which go back to Roman law, this paper argues that the current state of the art individual fairness corresponds most closely to the legal and technical circumstances. In contrast, group fairness cannot meet the legal requirements, since it is not possible to focus on the individual and to include specific circumstances of a person in the assessment of whether a fair decision has been made. Nevertheless, the analysis of the state-of-the-art methods for implementing fairness, or bias mitigation methods in AI systems shows a distinct focus on group fairness. This might be due to the somewhat agreed-upon definition for this term, as well as the easier implementation and evaluation of metrics and bias mitigation solutions. For this reason, group fairness is more in line with the intents of policymakers. It is also shown that inherent biases and discrimination are currently only insufficiently solved and are unintentionally perpetuated by the current developments of AI systems, leading to adverse consequences for certain demographic groups. Ensuring fairness in AI systems requires a multifaceted approach, ranging from interdisciplinary research in different fields such as law, computer science or philosophy, to deliberation, specific planning and evaluation of every aspect of development and use, including data collection, algorithm design and ongoing monitoring. By including diverse and representative datasets during the development phase, biases can be mitigated and the fairness of AI systems can be enhanced. Our paper contributes by providing a needed exchange between multiple disciplines to dissect and address the nuanced dimensions of fairness in AI, bridging the gap between technical development and legal considerations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.
[2] Mahoney T, Varshney K, Hind M. AI Fairness. O'Reilly Media, Incorporated; 2020.
[3] Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. 2023, http://dx.doi.org/10.48550/arXiv.2304.07683, arXiv.
[4] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilović A, Nagar S, Ramamurthy KN, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019;63(4/5):4:1–4:15. http://dx.doi.org/10.1147/JRD.2019.2942287.
[5] New Vantage Partners. Data and AI leadership executive survey 2022. 2022, https://www.newvantage.com/_files/ugd/e5361a_ad5a8b3da8254a71807d2dccdb0844be.pdf, [Accessed 07 April 2024].
[6] Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. Calif Manage Rev 2019;61(4):5–14. http://dx.doi.org/10.1177/0008125619864925.
[7] Zhang C, Lu Y. Study on artificial intelligence: The state of the art and future prospects. J Ind Inf Integr 2021;23:100224. http://dx.doi.org/10.1016/j.jii.2021.100224.
[8] Parnas DL. The real risks of artificial intelligence. Commun ACM 2017;60(10):27–31. http://dx.doi.org/10.1145/3132724.
[9] Blomberg T, Bales W, Mann K, Meldrum R, Nedelec J. Validation of the COMPAS risk assessment classification instrument. Tallahassee, FL: College of Criminology and Criminal Justice, Florida State University; 2010.
[10] Hamilton M. The sexist algorithm. Behav Sci Law 2019;37(2):145–57. http://dx.doi.org/10.1002/bsl.2406.
[11] Feuerriegel S, Dolata M, Schwabe G. Fair AI - Challenges and Opportunities. Bus Inf Syst Eng 2020;62:379–84. http://dx.doi.org/10.1007/s12599-020-00650-3.
[12] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 2012;33:1–33. http://dx.doi.org/10.1007/s10115-011-0463-8.
[13] Wan M, Zha D, Liu N, Zou N. In-processing modeling techniques for machine learning fairness: A survey. ACM Trans Knowl Discov Data 2023;17(3). http://dx.doi.org/10.1145/3551390.
[14] Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press; 2000.
[15] Schölkopf B. Causality for machine learning. In: Probabilistic and causal inference: the works of judea pearl. first ed.. New York, NY, USA: Association for Computing Machinery; 2022, p. 765–804. http://dx.doi.org/10.1145/3501714.3501755.
[16] Scherer MU. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harv J Law Technol 2015;29:353–400.
[17] Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv 2018;4(1). http://dx.doi.org/10.1126/sciadv.aao5580.

[18] Silvernail KD. Cross-cultural organizational justice: when are fairness perceptions universal or culturally dependent? (Ph.D. thesis), University of Massachusetts Amherst; 2016.

[19] James K. Culture and organizational justice: State of the literature and suggestions for future directions. In: Cropanzano RS, Ambrose ML, editors. The oxford handbook of justice in the workplace. Oxford, United Kingdom: Oxford University Press; 2015, p. 273–90.

[20] European Commission. Artificial intelligence, real benefits. 2018, https://digital-strategy.ec.europa.eu/en/library/artificial-intelligence-real-benefits, [Accessed 07 April 2024].

[21] Shin D, Park YJ. Role of fairness, accountability, and transparency in algorithmic affordance. Comput Hum Behav 2019;98:277–84. http://dx.doi.org/10.1016/j.chb.2019.04.019.

[22] Kim H, Shin S, Jang J, Song K, Joo W, Kang W, Moon I-C. Counterfactual fairness with disentangled causal effect variational autoencoder. Proce AAAI Conf Artif Intell 2021;35(9):8128–36. http://dx.doi.org/10.1609/aaai.v35i9.16990.

[23] Chouldechova A, Roth A. The frontiers of fairness in machine learning. 2018, http://dx.doi.org/10.48550/arXiv.1810.08810, arXiv.

[24] Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness. 2023, http://dx.doi.org/10.48550/arXiv.1808.00023, arXiv.

[25] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. 2016, http://dx.doi.org/10.48550/arXiv.1609.05807, arXiv.

[26] Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the international workshop on software fairness. FairWare '18, New York, NY, USA: Association for Computing Machinery; 2018, p. 1–7. http://dx.doi.org/10.1145/3194770.3194776.

[27] Minow M. EQUALITY VS. EQUITY. Am J Law Equal 2021;1:167–93. http://dx.doi.org/10.1162/ajle_a_00019.

[28] Tyler T. Procedural justice and the courts. Court Rev J Am Judges Assoc 2007;44(1/2):26–31.

[29] Rodrigues R. Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. J Responsib Technol 2020;4:100005. http://dx.doi.org/10.1016/j.jrt.2020.100005.

[30] Colquitt JA. On the dimensionality of organizational justice: A construct validation of a measure. J Appl Psychol 2001;86(3):386–400. http://dx.doi.org/10.1037/0021-9010.86.3.386.

[31] Wiseman J, Stillwell A. Organizational justice: Typology, antecedents and consequences. Encyclopedia 2022;2(3):1287–95. http://dx.doi.org/10.3390/encyclopedia2030086.

[32] Tyler T, Jackson J, Bradford B. Procedural justice and cooperation. In: Bruinsma G, Weisburd D, editors. Encyclopedia of criminology and criminal justice. New York, NY: Springer; 2014, p. 4011–24. http://dx.doi.org/10.1007/978-1-4614-5690-2_64.

[33] Rawls J. A Theory of Justice. Cambridge: Belknap Press of Harvard University Press; 1971.

[34] Lind A, Tyler T. The social psychology of procedural justice. Springer New York, NY; 1988, http://dx.doi.org/10.1007/978-1-4899-2115-4.

[35] Tyler T, Lind A. A relational model of authority in groups. Adv Exp Soc Psychol 1992;25:115–95.

[36] Goldman B, Cropanzano R. "Justice" and "fairness" are not the same thing. J Organ Behav 2015;36(2):313–8. http://dx.doi.org/10.1002/job.1956.

[37] Smith SA. In defence of substantive fairness. Law Q Rev 1996;112(1):138–58.

[38] Johnson JJ, Brunet E. Substantive fairness in securities arbitration. University Cincinnati Law Rev 2007;76:1–53.

[39] Buckley FH. Three theories of substantive fairness. Hofstra Law Rev 1990;19(1):33–66.

[40] Gentile G. Two strings to one bow? Article 47 of the EU charter of fundamental rights in the EU competition case law: Between procedural and substantive fairness. Mark Competition Law Rev 2020;4(2):169–204.

[41] Council of Europe. Convention for the Protection of Human Rights and Fundamental Freedoms. 1950, Council of Europe Treaty Series 005.

[42] Rozakis C. The right to a fair trial in civil cases. Judic Stud Inst J 2004;4(2):96–106.

[43] Mahoney P. Right to a fair trial in criminal matters under 107 article 6 e.c.h.r. Judic Stud Inst J 2004;4(2):107–29.

[44] European Convention. Charter of Fundamental Rights of the European Union. 2000, OJ C 364/1.

[45] Gutman K. The essence of the fundamental right to an effective remedy and to a fair trial in the case-law of the court of justice of the European union: The best is yet to come?. German Law J 2019;20(6):884–903. http://dx.doi.org/10.1017/glj.2019.67.

[46] Schwartz DS. Mandatory arbitration and fairness. Notre Dame Law Rev 2009;84(3):1247–342.

[47] American Arbitration Association. The code of ethics for arbitrators in commercial disputes. 2004, https://www.adr.org/sites/default/files/document_repository/Commercial_Code_of_Ethics_for_Arbitrators_2010_10_14.pdf.

[48] Centre VIA. VIAC rules of arbitration and mediation 2021. 2021.

[49] Union E. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). 2023, OJ L.

[50] European Union. Council directive 93/13/EEC of 5 april 1993 on unfair terms in consumer contracts. 1993, OJ L 95/29.

[51] European Union. Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive') (Text with EEA relevance). 2005, OJ L 149/22.

[52] OECD. Recommendation of the Council on Artificial Intelligence, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, [Accessed 07 April 2024].

[53] European Parliament. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (AI) and amending certain Union Legislative Acts. 2024, COM(2021)0206 – C9-0146/2021 – 2021/0106(COD).

[54] John-Mathews J, Cardon D, Balagué C. From reality to world. a critical perspective on AI fairness. J Bus Ethics 2022;178:945—-959. http://dx.doi.org/10.1007/s10551-022-05055-8.

[55] Dymitruk M. The right to a fair trial in automated civil proceedings. Masaryk Univ J Law Technol 2019;13(1):27–44. http://dx.doi.org/10.5817/MUJLT2019-1-2.

[56] Williams BA, Brooks CF, Shmargad Y. How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. J Inf Policy 2018;8:78–115. http://dx.doi.org/10.5325/jinfopoli.8.2018.0078.

[57] Lütz F. Gender equality and artificial intelligence in Europe. Addressing direct and indirect impacts of algorithms on gender-based discrimination. ERA Forum 2022;23:33–52. http://dx.doi.org/10.1007/s12027-022-00709-6.

[58] Wachter S. The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. 2022, http://dx.doi.org/10.48550/arXiv.2205.01166, arXix.

[59] European Union. Treaty on the Functioning of the European Union. 2012, OJ C 326/47.

[60] European Council. Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. 2000, OJ L 180/22.

[61] European Council. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. 2000, OJ L 303/16.

[62] Maliszewska-Nienartowicz J. Direct and indirect discrimination in European union law – how to draw a dividing line. Int J Soc Sci 2014;3(1):41–55.

[63] Prince AER, Schwarcz D. Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Rev 2020;105(3):1257–318.

[64] Hoffmann AL. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf Commun Soc 2019;22(7):900–15. http://dx.doi.org/10.1080/1369118X.2019.1573912.

[65] European Commission. A Union of Equality: Gender Equality Strategy 2020–2025. 2020, COM (2020) 152 final.

[66] European Commission. Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021, COM (2021) 206 final.

[67] European Union. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 2023, COM(2021)0206 – C9-0146/2021 – 2021/0106(COD).

[68] European Court of Justice. Case C-443/15, david l. Parris v trinity college dublin and others. 2016, EU:C:2016:897.

[69] Xenidis R. Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. Maastricht J Eur Comp Law 2020;27(6):736–58. http://dx.doi.org/10.1177/1023263X20982173.

[70] Heinrichs B. Discrimination in the age of artifcial intelligence. AI & SOCIETY 2022;37:143–54. http://dx.doi.org/10.1007/s00146-021-01192-2.

[71] Wairimu N. Dignity as non-discrimination: Existential protests and legal claim-making for reproductive rights. Phil. Soc Crit 2017;43(1):51–82. http://dx.doi.org/10.1177/0191453716645145.

[72] Weinberg L. Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. J Artificial Intelligence Res 2022;74:75–109.

[73] Nielsen A. Practical fairness. O'Reilly Media, Inc., Sebastopol; 2020.

[74] Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Annu Rev Stat Appl 2021;8(1):141–63. http://dx.doi.org/10.1146/annurev-statistics-042720-125902.

[75] Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2022;55(3):1–44. http://dx.doi.org/10.1145/3494672.

[76] Hutchinson B, Mitchell M. 50 Years of test (un)fairness: Lessons for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. FAT* '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 49–58. http://dx.doi.org/10.1145/3287560.3287600.

[77] Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints. In: 2009 IEEE international conference on data mining workshops. 2009, p. 13–8. http://dx.doi.org/10.1109/ICDMW.2009.83.

[78] Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. ITCS '12, New York, NY, USA: Association for Computing Machinery; 2012, p. 214–26. http://dx.doi.org/10.1145/2090236.2090255.

[79] Žliobaitė I. Measuring discrimination in algorithmic decision making. Data Min Knowl Discov 2017;31:1060–89. http://dx.doi.org/10.1007/s10618-017-0506-1.

[80] Lohia PK, Natesan Ramamurthy K, Bhide M, Saha D, Varshney KR, Puri R. Bias mitigation post-processing for individual and group fairness. In: ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing. ICASSP, 2019, p. 2847–51. http://dx.doi.org/10.1109/ICASSP.2019.8682620.

[81] Kearns M, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Dy J, Krause A, editors. Proceedings of the 35th international conference on machine learning. Proceedings of machine learning research, 80, Stockholmsmässan, Stockholm, Sweden: PMLR; 2018, p. 2564–72, URL https://proceedings.mlr.press/v80/kearns18a.html.

[82] Binns R. On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT* '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 514–24. http://dx.doi.org/10.1145/3351095.3372864.

[83] Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: The state of the art. Sociol Methods Res 2021;50(1):3–44. http://dx.doi.org/10.1177/0049124118782533.

[84] Wu B, Han K, Zhang E. On the task assignment with group fairness for spatial crowdsourcing. Inf Process Manage 2023;60:103175. http://dx.doi.org/10.1016/j.ipm.2022.103175.

[85] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C, editors. Proceedings of the 1st conference on fairness, accountability and transparency. Proceedings of machine learning research, 81, New York, NY, USA: PMLR; 2018, p. 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html.

[86] Garg P, Villasenor J, Foggo V. Fairness metrics: A comparative analysis. In: 2020 IEEE international conference on big data (big data). 2020, p. 3662–6. http://dx.doi.org/10.1109/BigData50022.2020.9378025.

[87] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of the 30th international conference on neural information processing systems. NIPS '16, Red Hook, NY, USA: Curran Associates Inc.; 2016, p. 3323–31.

[88] Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. 2017, http://dx.doi.org/10.48550/arXiv.1702.06081, arXiv.

[89] Jo N, Tang B, Dullerud K, Aghaei S, Rice E, Vayanos P. Fairness in contextual resource allocation systems: Metrics and incompatibility results. In: Proceedings of the AAAI conference on artificial intelligence. 37, (10):2023, p. 11837–46. http://dx.doi.org/10.1609/aaai.v37i10.26397.

[90] Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. In: Proceedings of the 31st international conference on neural information processing systems. NIPS '17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 5684–93.

[91] Diana E, Gill W, Kearns M, Kenthapadi K, Roth A. Minimax group fairness: Algorithms and experiments. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society. AIES '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 66–76. http://dx.doi.org/10.1145/3461702.3462523.

[92] Galhotra S, Brun Y, Meliou A. Fairness testing: Testing software for discrimination. In: Proceedings of the 2017 11th joint meeting on foundations of software engineering. ESEC/FSE 2017, New York, NY, USA: Association for Computing Machinery; 2017, p. 498–510. http://dx.doi.org/10.1145/3106237.3106277.

[93] Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Proceedings of the 31st international conference on neural information processing systems. NIPS '17, 30, Curran Associates Inc.; 2017, p. 4066–76.

[94] Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. The case for process fairness in learning: Feature selection for fair decision making. In: symposium on machine learning and the law at the 29th conference on neural information processing systems. 1, (2):2016, p. 1–11.

[95] Ingold D, Soper S. Amazon doesn't consider the race of its customers. Should it? 2016, URL https://www.bloomberg.com/graphics/2016-amazon-same-day/, [Accessed 07 April 2024].

[96] Karimi H, Akbar Khan MF, Liu H, Derr T, Liu H. Enhancing individual fairness through propensity score matching. In: 2022 IEEE 9th international conference on data science and advanced analytics. DSAA, Shenzhen, China; 2022, p. 1–10. http://dx.doi.org/10.1109/DSAA54385.2022.10032333.

[97] Dutta S, Wei D, Yueksel H, Chen P-Y, Liu S, Varshney KR. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In: Proceedings of the 37th international conference on machine learning. ICML '20, JMLR.org; 2020, p. 2803–13.

[98] Adams-Prassl J, Binns R, Kelly-Lyth A. Directly discriminatory algorithms. Mod Law Rev 2023;86(1):144–75. http://dx.doi.org/10.1111/1468-2230.12759.

[99] Nabi R, Shpitser I. Fair inference on outcomes. In: Proceedings of the AAAI conference on artificial intelligence. AAAI'18/iAAI'18/eAAI'18, AAAI Press; 2018, p. 1931–40.

[100] Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B. Avoiding discrimination through causal reasoning. In: Proceedings of the 31st international conference on neural information processing systems. NIPS '17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 656–66.

[101] Jackson F. Stay-at-home dad's 'life ruined' by google after he was investigated when photos he took of his sick toddler son's genitals for a doctor to review during the pandemic were flagged by AI as potential child sexual abuse material. 2022, URL https://www.dailymail.co.uk/sciencetech/article-11133805/Google-AI-flags-dad-photos-childs-groin-infection-phone-share-doctors.html, [Accessed 07 April 2024].

[102] Mousourakis G. Roman Law and the Origins of the Civil Law Tradition. Cham: Springer International Publishing; 2015, http://dx.doi.org/10.1007/978-3-319-12268-7.

[103] Burdick WL. The Principles of Roman Law and Their Relation to Modern Law. Rochester, New York: The Lawbook Exchange, Ltd.; 2004.

[104] Trakman L. Ex aequo et bono: Demystifying an ancient concept. Chic J Int Law 2008;8(2):621–42.

[105] Lipton ZC, Chouldechova A, McAuley J. Does mitigating ml's impact disparity require treatment disparity? In: Proceedings of the 32nd international conference on neural information processing systems. NIPS '18, Red Hook, NY, USA: Curran Associates Inc.; 2018, p. 8136–46.

[106] Olteanu A, Castillo C, Diaz F, Kı cıman E. Social data: Biases, methodological pitfalls, and ethical boundaries. Front Big Data 2019;2. http://dx.doi.org/10.3389/fdata.2019.00013.

[107] Ruths D, Pfeffer J. Social media for large studies of behavior. Science 2014;346(6213):1063–4. http://dx.doi.org/10.1126/science.346.6213.1063.

[108] Hellström T, Dignum V, Bensch S. Bias in machine learning – what is it good for? 2020, http://dx.doi.org/10.48550/arXiv.2004.00686, arXiv.

[109] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6):1–35. http://dx.doi.org/10.1145/3457607.

[110] Fahse T, Huber V, van Giffen B. Managing bias in machine learning projects. In: Ahlemann F, Schütte R, Stieglitz S, editors. Innovation through information systems. Cham: Springer International Publishing; 2021, p. 94–109. http://dx.doi.org/10.1007/978-3-030-86797-3_7.

[111] Bennett B, Cohn AG. Automated common-sense spatial reasoning: Still a huge challenge. In: Muggleton S, Chater N, editors. Human-like machine intelligence. Oxford University Press; 2021, p. 405–29. http://dx.doi.org/10.1093/oso/9780198862536.003.0020.

[112] Jin Z, Liu J, Zhiheng L, Poff S, Sachan M, Mihalcea R, Diab MT, Schölkopf B. Can large language models infer causation from correlation? In: The twelfth international conference on learning representations. 2023.

[113] Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and access in algorithms, mechanisms, and optimization. EAAMO '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 1–9. http://dx.doi.org/10.1145/3465416.3483305.

[114] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst 2023;3:121–54. http://dx.doi.org/10.1016/j.iotcps.2023.04.003.

[115] Abdollahpouri H, Mansoury M, Burke R, Mobasher B. The unfairness of popularity bias in recommendation. 2019, http://dx.doi.org/10.48550/arXiv.1907.13286, arXiv.

[116] Freire M, de Castro L. E-recruitment recommender systems: a systematic review.. Knowl Inf Syst 2021;63:1–20. http://dx.doi.org/10.1007/s10115-020-01522-8.

[117] Unkelbach C, Alves H, Koch A. Chapter three - negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In: Gawronski B, editor. Advances in experimental social psychology. vol. 62, Academic Press; 2020, p. 115–87. http://dx.doi.org/10.1016/bs.aesp.2020.04.005.

[118] Straw I. The automation of bias in medical artificial intelligence (AI): Decoding the past to create a better future. Artif Intell Med 2020;110:101965. http://dx.doi.org/10.1016/j.artmed.2020.101965.

[119] Korteling JEH, van de Boer-Visschedijk G, Blankendaal RAM, Boonekamp R, Eikelboom A. Human- versus artificial intelligence. Frontiers in Artificial Intelligence 2021;4. http://dx.doi.org/10.3389/frai.2021.622364.

[120] Skitka LJ, Mosier K, Burdick M. Does automation bias decision-making?. Int J Hum-Comput Stud 1999;51(5):991–1006. http://dx.doi.org/10.1006/ijhc.1999.0252.

[121] Mökander J, Morley J, Taddeo M, Floridi L. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. Sci Eng Ethics 2021;27. http://dx.doi.org/10.1007/s11948-021-00319-4.
[122] Coca-Vila I. Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. Crim Law Philos 2018;12:59–82. http://dx.doi.org/10.1007/s11572-017-9411-3.
[123] Lü L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, Zhou T. Recommender systems. Phys Rep 2012;519(1):1–49. http://dx.doi.org/10.1016/j.physrep.2012.02.006.
[124] Babic B, Gerke S, Evgeniou T, Cohen IG. Direct-to-consumer medical machine learning and artificial intelligence applications. Nat Mach Intell 2021;3:283–7. http://dx.doi.org/10.1038/s42256-021-00331-0.
[125] John T. Israel is using artificial intelligence to help pick bombing targets in gaza, report says. 2024, URL Accessed07April2024,
[126] Skitka LJ, Mosier K, Burdick MD. Accountability and automation bias. Int J Hum-Comput Stud 2000;52(4):701–17. http://dx.doi.org/10.1006/ijhc.1999.0349.
[127] Mosier KL, Skitka LJ. Automation use and automation bias. Proc Hum Factors Ergon Soc Annu Meet 1999;43(3):344–8. http://dx.doi.org/10.1177/154193129904300346.
[128] O'neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, New York; 2016.
[129] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big?. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. FAccT '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 610–23. http://dx.doi.org/10.1145/3442188.3445922.
[130] Hooker S. Moving beyond "algorithmic bias is a data problem". Patterns 2021;2(4):100241. http://dx.doi.org/10.1016/j.patter.2021.100241.
[131] Jiang H, Nachum O. Identifying and correcting label bias in machine learning. In: Chiappa S, Calandra R, editors. Proceedings of the twenty third international conference on artificial intelligence and statistics. Proceedings of machine learning research, 108, Palermo, Italy: PMLR; 2020, p. 702–12, URL https://proceedings.mlr.press/v108/jiang20a.html.
[132] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, Hicklen RS, Moukheiber L, Moukheiber D, Ma H, Mathur P. Bias in artificial intelligence algorithms and recommendations for mitigation.. PLOS Digital Health 2023;2(6):1–14. http://dx.doi.org/10.1371/journal.pdig.0000278.
[133] Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018, p. 15–20. http://dx.doi.org/10.18653/v1/N18-2003.
[134] Cotter A, Jiang H, Sridharan K. Two-player games for efficient non-convex constrained optimization. In: Garivier A, Kale S, editors. Proceedings of the 30th international conference on algorithmic learning theory. Proceedings of machine learning research, 98, PMLR; 2019, p. 300–32, URL https://proceedings.mlr.press/v98/cotter19a.html.
[135] Cotter A, Jiang H, Gupta MR, Wang SL, Narayan T, You S, Sridharan K. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals.. J Mach Learn Res 2019;20(172):1–59.
[136] Hussain H, Cao M, Sikdar S, Helic D, Lex E, Strohmaier M, Kern R. Adversarial inter-group link injection degrades the fairness of graph neural networks. In: 2022 IEEE international conference on data mining. ICDM, 2022, p. 975–80. http://dx.doi.org/10.1109/ICDM54844.2022.00117.
[137] Hort M, Chen Z, Zhang JM, Sarro F, Harman M. Bias mitigation for machine learning classifiers: A comprehensive survey. 2022, http://dx.doi.org/10.48550/arXiv.2207.07068, arXiv.
[138] Feldman T, Peake A. End-to-end bias mitigation: Removing gender bias in deep learning. 2023, http://dx.doi.org/10.48550/arXiv.2104.02532, arXiv.
[139] Yurochkin M, Bower A, Sun Y. Training individually fair ML models with sensitive subspace robustness. 2020, http://dx.doi.org/10.48550/arXiv.1907.00020, arXiv.
[140] Dwork C, Immorlica N, Kalai AT, Leiserson M. Decoupled classifiers for group-fair and efficient machine learning. In: Friedler SA, Wilson C, editors. Proceedings of the 1st conference on fairness, accountability and transparency. Proceedings of machine learning research, 81, PMLR; 2018, p. 119–33, URL https://proceedings.mlr.press/v81/dwork18a.html.
[141] Ustun B, Liu Y, Parkes D. Fairness without harm: Decoupled classifiers with preference guarantees. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, 97, PMLR; 2019, p. 6373–82.
[142] Suriyakumar VM, Ghassemi M, Ustun B. When personalization harms performance: Reconsidering the use of group attributes in prediction. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th international conference on machine learning. Proceedings of machine learning research, 202, PMLR; 2023, p. 33209–28.
[143] Liu S, Vicente LN. Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. Comput Manag Sci 2022;19:513–37. http://dx.doi.org/10.1007/s10287-022-00425-z.
[144] Bank D, Koenigstein N, Giryes R. Autoencoders. In: Rokach L, Maimon O, Shmueli E, editors. Machine learning for data science handbook: data mining and knowledge discovery handbook. Cham: Springer International Publishing; 2023, p. 353–74. http://dx.doi.org/10.1007/978-3-031-24628-9_16.
[145] Borges R, Stefanidis K. On mitigating popularity bias in recommendations via variational autoencoders. In: Proceedings of the 36th annual ACM symposium on applied computing. SAC '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 1383–9. http://dx.doi.org/10.1145/3412841.3442123.
[146] Liang D, Krishnan RG, Hoffman MD, Jebara T. Variational autoencoders for collaborative filtering. In: Proceedings of the 2018 world wide web conference. WWW '18, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2018, p. 689—698. http://dx.doi.org/10.1145/3178876.3186150.
[147] Grari V, Lamprier S, Detyniecki M. Fairness without the sensitive attribute via causal variational autoencoder. In: Raedt LD, editor. Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization; 2022, p. 696–702. http://dx.doi.org/10.24963/ijcai.2022/98.
[148] Tyagi S, Xie J, Andrews R. E-VAN: Enhanced variational AutoEncoder network for mitigating gender bias in static word embeddings. In: Proceedings of the 2022 6th international conference on natural language processing and information retrieval. NLPIR '22, New York, NY, USA: Association for Computing Machinery; 2023, p. 57–64. http://dx.doi.org/10.1145/3582768.3582804.
[149] Hort M, Zhang JM, Sarro F, Harman M. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In: Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. ESEC/FSE 2021, New York, NY, USA: Association for Computing Machinery; 2021, p. 994–1006. http://dx.doi.org/10.1145/3468264.3468565.
[150] Currie GM. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?. Semin Nucl Med 2023;53(5):719–30. http://dx.doi.org/10.1053/j.semnuclmed.2023.04.008.
[151] Bansal G, Hosack B, Iversen J, Mitchell A, Hadidi R, George JF. ChatGPT – another hype or out-of-this-world?. Journal of the Midwest Association for Information Systems (JMWAIS) 2023;2023(2):29–36.
[152] Hepp A, Loosen W, Dreyer S, Jarke J, Kannengieß er S, Katzenbach C, Malaka R, Pfadenhauer M, Puschmann C, Schulz W. ChatGPT, lamda, and the hype around communicative AI: The automation of communication as a field of research in media and communication studies. Human-Machine Commun 2023;6:41–63. http://dx.doi.org/10.30658/hmc.6.4.
[153] O'Leary DE. An analysis of watson vs. BARD vs. ChatGPT: The jeopardy! challenge. AI Mag 2023;44(3):282–95. http://dx.doi.org/10.1002/aaai.12118.
[154] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. Llama 2: Open foundation and fine-tuned chat models. 2023, http://dx.doi.org/10.48550/arXiv.2307.09288, arXiv.
[155] McGee RW. ChatGPT and copyright infringement: An exploratory study. ResearchGate 2023. http://dx.doi.org/10.13140/RG.2.2.35449.03684.
[156] Dave T, Athaluri SA, Singh S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. http://dx.doi.org/10.3389/frai.2023.1169595.
[157] Cotton RE, Debby PAC, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int 2023;1–12. http://dx.doi.org/10.1080/14703297.2023.2190148.
[158] Garante per la protezione dei dati personali. Artificial intelligence: stop to ChatGPT by the Italian SA personal data is collected unlawfully, no age verification system is in place for children. 2023, URL Accessed07April2024.
[159] Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue J, Wu Y. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. 2023, http://dx.doi.org/10.48550/arXiv.2301.07597, arXiv.
[160] Sousa S, Kern R. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. Artif Intell Rev 2023;56(2):1427–92. http://dx.doi.org/10.1007/s10462-022-10204-6.
[161] Rozado D. The political biases of ChatGPT. Soc Sci 2023;12(3). http://dx.doi.org/10.3390/socsci12030148.
[162] 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). 2016, OJ L 119/1.
[163] Kutyłowski M, Lauks-Dutka A, Yung M. GDPR – challenges for reconciling legal rules with technical reality. In: Chen L, Li N, Liang K, Schneider S, editors. Computer security – ESORICS 2020. Lecture notes in computer science, 12308, Cham: Springer; 2020, http://dx.doi.org/10.1007/978-3-030-58951-6_36.

[164] Häuselmann A, Custers B. Substantive fairness in the GDPR: Fairness elements for article 5.1a GDPR. Comput Law Secur Rev 2024;52:105942. http://dx.doi.org/10.1016/j.clsr.2024.105942.
[165] European Commission. Ethics Guidelines for Trustworthy AI. 2019, http://dx.doi.org/10.2759/346720.
[166] Tschantz MC. What is proxy discrimination? In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. FAccT '22, New York, NY, USA: Association for Computing Machinery; 2022, p. 1993–2003. http://dx.doi.org/10.1145/3531146.3533242.
[167] Martínez M, Segura F, Andújar JM, Ceada Y. The gender gap in STEM careers: An inter-regional and transgenerational experimental study to identify the low presence of women. Educ Sci 2023;13(7):649. http://dx.doi.org/10.3390/educsci13070649.
[168] Hauer MP, Kevekordes J, Haeri MA. Legal perspective on possible fairness measures – a legal discussion using the example of hiring decisions. Comput Law Secur Rev 2021;42:105583. http://dx.doi.org/10.1016/j.clsr.2021.105583.
[169] Kalimo H, Majcher K. The concept of fairness: Linking EU competition and data protection law in the digital marketplace. Eur Law Rev 2017;47(2):210–33.
[170] Landers RN, Behrend TS. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. Am Psychol 2023;78(1):36–49. http://dx.doi.org/10.1037/amp0000972.
[171] Tsagourias N. Digitalization and its systemic impact on the use of force regime: Legal uncertainty and the replacement of international law. Ger Law J 2023;24(3):494–507. http://dx.doi.org/10.1017/glj.2023.33.
[172] Gal MS, Aviv O. The competitive effects of the GDPR. J Compet Law Econ 2020;16(3):349–91. http://dx.doi.org/10.1093/joclec/nhaa012.
[173] Königstorfer F, Thalmann S. AI documentation: A path to accountability. J Responsib Technol 2022;11:100043. http://dx.doi.org/10.1016/j.jrt.2022.100043.
[174] Wachter S, Mittelstadt B, Russell C. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 2021;41:105567. http://dx.doi.org/10.1016/j.clsr.2021.105567.
[175] Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B. Trustworthy AI: From principles to practices. ACM Comput Surv 2023;55(9):1–46. http://dx.doi.org/10.1145/3555803.
[176] Varona D, Suárez JL. Discrimination, bias, fairness, and trustworthy AI. Appl Sci 2022;12(12):5826. http://dx.doi.org/10.3390/app12125826.
[177] Shin D, Shin EY. Data's impact on algorithmic bias. Computer 2023;56(6):90–4. http://dx.doi.org/10.1109/MC.2023.3262909.
[178] Demelius L, Kern R, Trügler A. Recent advances of differential privacy in centralized deep learning: A systematic survey. 2023, http://dx.doi.org/10.48550/arXiv.2309.16398, arXiv.
[179] Häuselmann A, Sears AM, Zard L, Fosch-Villaronga E. EU law and emotion data. In: 11th international conference on affective computing and intelligent interaction. ACII, IEEE Computer Society; 2023, p. 1–8. http://dx.doi.org/10.1109/ACII59096.2023.10388181.
[180] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. NIPS '17, Red Hook, NY, USA: Curran Associates Inc.; 2017, p. 4768–77.
[181] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: Association for Computing Machinery; 2016, p. 1135–44. http://dx.doi.org/10.1145/2939672.2939778.
[182] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT* '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 607–17. http://dx.doi.org/10.1145/3351095.3372850.
[183] Xu W. Toward human-centered AI: A perspective from human-computer interaction. Interactions 2019;26(4):42–6. http://dx.doi.org/10.1145/3328485.
[184] Smuha NA. From a 'race to ai' to a 'race to AI regulation': regulatory competition for artificial intelligence. Law Innov Technol 2021;13(1):57–84. http://dx.doi.org/10.1080/17579961.2021.1898300.
[185] Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. FAccT '23, New York, NY, USA: Association for Computing Machinery; 2023, p. 1112–23. http://dx.doi.org/10.1145/3593013.3594067.
[186] Cerutti F, Kaplan LM, Kimmig A, Şensoy M. Handling epistemic and aleatory uncertainties in probabilistic circuits. Mach Learn 2022;111:1259–301. http://dx.doi.org/10.1007/s10994-021-06086-4.
[187] Undheim K, Erikson T, Timmermans B. True uncertainty and ethical AI: Regulatory sandboxes as a policy tool for moral imagination. AI Ethics 2023;3:997–1002. http://dx.doi.org/10.1007/s43681-022-00240-x.
[188] Liu A, Guerra S, Fung I, Matute G, Kamar E, Lasecki W. Towards hybrid human-AI workflows for unknown unknown detection. In: Proceedings of the web conference 2020. WWW '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 2432–42. http://dx.doi.org/10.1145/3366423.3380306.