It's complicated. The relationship of algorithmic fairness and non-discrimination regulations for high-risk systems in the EU AI Act

Anonymous Author(s)
Affiliation
Address
email

Abstract

What constitutes a fair decision? This question is not only difficult to answer for humans but becomes more challenging when Artificial Intelligence (AI) models are used. In light of problematic algorithmic outcomes, the EU has recently passed the AI Act, which mandates specific rules for high-risk systems, incorporating both traditional legal non-discrimination regulations and machine learning based algorithmic fairness concepts. This paper aims to bridge these two concepts in the AI Act by providing: (1) a high-level introduction targeting computer science-oriented scholars, and (2) an analysis of the relationship between the AI Act's legal non-discrimination regulations and its algorithmic fairness provisions. Finally, we consider future steps in the application of non-discrimination regulations and the AI Act regulations. This paper serves as a foundation for future interdisciplinary collaboration between legal scholars and machine learning researchers with a computer science background studying discrimination in AI systems.

1 Introduction

How can we ensure fair algorithms in the context of AI systems? When regulating algorithms and, more specifically, Artificial Intelligence (AI) systems, this question becomes fundamental. While the question is also crucial in a non-digital world, it becomes increasingly pressing in the digital world. The European Union (EU) has become one of the forerunners in regulating the digital age. In order to maintain responsible data processing in the machine learning domain, the European Union (EU) recently adopted the Artificial Intelligence Act (AIA)[1]. The aim of this AI regulation is to maintain a level playing field for "ethical principles" [Nolte et al., 2024] within and outside the EU (see recital 27 AIA).

Fair algorithmic processing also matters from a computational perspective. Over the past decade, algorithmic fairness has become a well-established field within machine learning that focuses on defining, mitigating, and evaluating the discriminatory behavior of Artificial Intelligence models [Pessach and Shmueli, 2023]. The importance becomes evident when looking at several discriminatory behaviours of algorithms in the past, ranging from hiring [Dastin, 2022] to social welfare systems [Hadwick and Lan, 2021]. In recent years, large generative (multimodal) models, particularly Large Language Models (LLMs) with their high accessibility and wide range of applications, have posed significant new challenges to the algorithmic fairness domain [Chu et al., 2024, Kotek et al., 2023].

[1] Regulation (EU) 2024/1689.

Submitted to Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1.1 Previous literature.

Discrimination through algorithms is not a recent phenomenon and was analysed prior to the AI Act. As early as the 1980s, algorithmic discrimination was observed in admissions settings [Connors et al., 1981, Williams et al., 1981], and gender inequalities in educational software were discussed [Huff and Cooper, 1987]. In 1996, researchers pointed out the levels at which such technical constraints and social institutions' biases can occur [Friedman and Nissenbaum, 1996]. More recently, the interaction between AI and automated decision-making systems with traditional European and US non-discrimination law has been studied extensively. [Barocas and Selbst, 2016] presented a taxonomy of different sources of discrimination and their impact on humans. The relationship among various European non-discrimination laws has also been studied [Wachter et al., 2021, 2020, Hacker, 2018, Weerts et al., 2023, Xenidis and Senden, 2020, Lewis et al., 2025]. For example, the seminal study by Wachter et al. [2021] presented a fairness metric that connects algorithmic fairness to the jurisprudence of the European Court of Justice in relation to the General Data Protection Regulation. Similarly, the connection between non-discrimination law and the GDPR has been analysed to "unlock the algorithmic black box" [Hacker, 2018]. Naturally, since these papers were published before the AI Act was passed, they do not contain references to it.

Recently, several papers have addressed the AI Act. Short studies such as [Deck et al., 2024a, Ruohonen, 2024] have analysed the relationship between fairness and the AI Act. Additionally, sections within papers of broader scope beyond pure non-discrimination analyses have mentioned aspects of the relationship between algorithmic fairness and the AI Act [Wachter, 2024, Hacker et al., 2024b, Novelli et al., 2024]. The AI Act and its relation to gender equality and non-discrimination law have also been discussed [Lütz, 2024]. Among these, the work by Bosoer et al. [2023] is most similar to ours in terms of its focus on the AI Act; however, it investigates non-discrimination regulations in the draft of the AI Act without a strong focus on the interaction with computer science. The most similar work in terms of the interaction between computer science and legal research is the paper by Weerts et al. [2023]. Unlike our work, theirs did not focus on the AI Act.

1.2 Our contributions.

This paper aims to foster and extend an interdisciplinary view on explicit and implicit fairness regulations in the AI Act. For the analysis of the EU AI Act's non-discrimination requirements in Section 2&3, our contributions are as follows:

1. We present the history, scope, and intentions of non-discrimination regulations in the AI Act.
2. We analysed the relation between high-risk systems' regulations and algorithmic fairness, finding that specific regulations will benefit from the forthcoming standardisation process.
3. We discuss future steps for the application of the regulations for algorithmic fairness in the context of classical non-discrimination regulations and the standardisation process at the intersection of the AI Act.

2 A primer on EU non-discrimination regulations

Interdisciplinary research on algorithmic fairness poses challenges for both computer scientists and legal scholars specialising in non-discrimination law. Understanding legal reasoning can be challenging without prior legal knowledge, just as understanding algorithmic methods is difficult without a computational background.

However, the complexity of interdisciplinary work goes beyond technical expertise. The challenges begin with terminology: For instance, in computer science, "fairness" is a term that can refer to different desiderata that aim to prevent socially or morally undesirable behavior or outcomes of algorithms. Although fairness has been discussed as a principle in economic law contexts [Scheuerer, 2023], it is not a specific legal term. The closest legal term is arguably "non-discrimination", which focuses on preventing unfair treatment based on characteristics such as race, gender, or other attributes on legal grounds. The interaction between these key concepts from law and computer science is demanding.

Therefore, we briefly introduce EU non-discrimination law for computer scientists in the following section[2]. We first describe fundamental rights as well as traditional EU non-discrimination law. Both influence the AI Act.

2.1 Fundamental rights in EU law.

The Charter of Fundamental Rights of the European Union (CFR) is the main fundamental rights regulation within EU law[3]. Under Article 6(1) of the Treaty on European Union, the Charter forms part of the primary law of the EU.

The CFR includes non-discrimination law. Article 21 CFR states: "Any discrimination based on any ground such as sex, race, color, 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 shall be prohibited." Thus, only discrimination based on these listed attributes is targeted.

However, the right to non-discrimination is not absolute. It is important to note that the right to non-discrimination is only one right in the CFR, among others, such as the freedom of expression and information (Article 11(1) CFR). Interpretation and application of the Charter require balancing rights as described in Article 52(1) CFR. First, legislators must balance different CFR rights proportionally when creating new laws. Additionally, when laws reference the CFR, implementers must also follow this balancing obligation. A logic of proportionality guides the assessment of fundamental rights [Almada and Petit, 2023]. Fundamental rights are neither absolute, hierarchical, nor quantifiable, and must be applied on a case-by-case basis [Sousa e Silva, 2024]. The provisions of the CFR apply to public parties (vertical applicability, Fornasier [2015]). Therefore, there are debates about how exactly the CFR applies to relationships between private parties (horizontal applicability) in general [Fornasier, 2015, Cherednychenko, 2007, Frantziou, 2015, Prechal, 2020] and in the AI Act context specifically [Lewis et al., 2025]. In the context of the GDPR, scholars and the European Court of Justice (ECJ) noted that the CFR can have a horizontal effect if the secondary law reflects a general principle of EU law [Ufert, 2020]. Since the AI Act also reflects general principles of EU law, we do not exclude some horizontal applicability of the CFR. When discussing fundamental rights in the AI Act, it is also essential to consider the broader political and institutional landscape [Palmiotto, 2025].

2.2 Direct and indirect discrimination in EU Law.

While the CFR is part of EU primary law, it also exerts influence over secondary EU law [Lewis et al., 2025]. EU law has a long history of non-discrimination regulations and different laws include non-discrimination regulations[4]. For example, the Directive 2000/43/EC (Race Equality Directive) and Directive 2000/78/EC (Employment Equality Directive) include specific non-discrimination regulations.

In order to legally assess discrimination under specific secondary EU law, two types of discrimination are differentiated: direct and indirect discrimination [Zuiderveen Borgesius et al., 2024, Wachter et al., 2020]. However, importantly, EU-based direct discrimination does not require any intentional wrongdoing [Weerts et al., 2023, Xenidis and Senden, 2020, Adams-Prassl et al., 2023].

Direct discrimination is defined as situations in which "one person is treated less favourably than another is, has been or would be treated in a comparable situation" (Article 2(2)(a) Directive 2000/43/EC). This means that discrimination occurs when individuals are treated less favorably on the basis of a protected attribute listed in the Article.

Indirect discrimination is more difficult to address. In indirect discrimination cases, seemingly neutral attributes are used, but they rely on a protected attribute. For indirect discrimination, it is important

[2]See section A.1 in the appendix for an introduction of algorithmic fairness to legal scholars.

[3]The CFR should not be confused with the European Convention on Human Rights (ECHR), which, though separate from the EU, is an agreement the EU is expected to accede to. The relationship between the ECHR and EU law is complex (see [Brittain, 2015]).

[4]See for an overview: `https://commission.europa.eu/aid-development-cooperation-fundamental-rights/your-fundamental-rights-eu/know-your-rights/equality/non-discrimination_en`.

to note that it can be justified through legitimate aims and appropriate means [Zuiderveen Borgesius et al., 2024].

Also, a third category of discrimination needs to be taken into account[5], which challenges traditional non-discrimination law: intersectional discrimination. Intersectional discrimination concerns cases "originating in several inextricably linked vectors of disadvantage" [Xenidis, 2023]. This becomes especially important since (modern) AI models do not use single variables as input but instead use many different aspects as input. This can lead to effects where discrimination may only occur at the intersection of gender and age [Weerts et al., 2023]. Whether, and to what extent, the ECJ currently recognises intersectional discrimination as a distinct form of discrimination remains an open question[Weerts et al., 2023, Xenidis, 2023, Atrey, 2018].

3 EU AI Act's non-discrimination regulations for high-risk systems.

Before we discuss the specific non-discrimination regulations for high-risk systems, we will briefly introduce the history, scope, and most important definitions of the AI Act. This short introduction is followed by an overview of the AI Act's non-discrimination regulations for high-risk systems.

3.1 The EU AI Act: History, Scope, and Definitions

In this section, we first describe the origins of the EU AI Act and clarify key concepts such as risk, systems, and the difference between developers and deployers, which is necessary to understand our analysis of the AI Act.

Emergent technology can benefit from efficient regulation. As one of the first comprehensive AI regulations worldwide [Wodi, 2024], the AI Act (AIA) introduces harmonised rules and obligations for the use and "placing on the market" of AI systems within the European Union (Article 1(2) AIA). It is primarily a product safety regulation [Almada and Petit, 2023], aiming to establish a level playing field for AI technologies across the Union [Nolte et al., 2024]. Notably, the original drafts of the EU Commission did not include explicit provisions on individual rights [Hacker et al., 2024b]. At this stage, individual rights — understood as protecting the rights of individuals affected — were largely absent. In fact, Members of the European Parliament initially did not prioritise the regulation of algorithmic discrimination [Chiappetta, 2023]. However, over the course of the legislative process, different aspects of non-discrimination regulations were integrated into the final text.

The AI Act relies in part on the New Legislative Framework of the European Union (EU) [Kaminski and Selbst, 2025]. The New Legislative Framework is a cornerstone of modern product safety law in the EU (see also [Commission, 2022] for further details), which also applies, for example, to medical devices and children's toys[6].

The AI Act regulates AI systems and also includes provisions for General Purpose AI (GPAI) models. The relationship or distinction between a model and a system in the context of the AIA remains legally unresolved. Article 3(1) defines an "AI system" as "a machine-based system that is designed to operate with varying levels of autonomy and 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", but the article does not provide a definition of "AI model". Recital 97 clarifies that "although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as, for example, a user interface, to become AI systems." This formulation does not conclusively settle the terminological relationship between models and systems, as discussed in more detail in [Nolte et al., 2024]. Moreover, the regulation of "AI systems" contrasts with the (trustworthy) computer science literature, which typically studies machine learning models. Since our focus is the legal framework, we follow the terminology of the AI Act, even though computer science literature tends to focus on models rather than systems.

[5]Please note that other non-discrimination rules in sector-specific regulations, such as the EU Consumer Credit Directive, are beyond the scope of our paper.

[6]See `https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en` for an overview.

The core regulatory structure of the AI Act is built on a risk-based approach [Hacker et al., 2024a]. Risk is defined in Article 3(2) AIA as "the combination of the probability of the occurrence of harm and the severity of that harm." Based on this definition, the AI Act categorises AI systems into different risk levels, each associated with a specific set of regulatory requirements[7]:

- Unacceptable risk (Art. 5 AIA): These AI systems are prohibited. Examples include social scoring by governments.
- High-risk (Art. 6ff. AIA): These systems are subject to stringent obligations, such as requirements for robustness, accuracy, or non-discrimination. Examples include AI used in recruitment, credit scoring, or law enforcement.
- Certain AI systems (Art. 50 AIA) with specific risk: These systems must meet transparency obligations, such as informing users they are interacting with an AI system. Examples include systems that produce deep-fakes.
- All other systems: These systems are not subject to specific regulatory obligations under the AIA, except broad regulations such as Article 4 AIA (AI Literacy).

In addition, the AIA includes specific provisions for General Purpose AI (GPAI) models (Articles 51 ff. AIA). The use case is not predefined, which is characteristic of GPAI models. LLMs are an example of a classical GPAI model. The AI Act distinguishes between general-purpose models posing "systemic risk" and those that do not.

The primary addressees of the AI Act are providers (Article 3(3) AIA) and deployers (Article 3(4) AIA). Providers are the entities responsible for developing and placing an AI model on the market or putting an AI system into service. In contrast, deployers are those who use an AI system under their authority. Many of the obligations we discuss in the following sections primarily concern the providers of AI systems, while some apply specifically to deployers (e.g., Article 26 AIA).

3.2 Non-discrimination regulations within the AI Act.

We began by scanning the AI Act for non-discrimination-related terms. In total, we scanned the AI Act for non-discrimination-related terms: discrimination, fundamental right, fairness, and bias. We noticed that within the definitions of Article 3 AIA, the terms serious incident (Article 3(49) AIA) and systemic risk (Article 3(65) AIA) refer to fundamental rights. Thus, these were included in our analysis as well. A full table with all articles of the AI Act, including non-discrimination-related terms, can be found in the appendix in Table 1.

This analysis shows, first, that the majority of non-discrimination regulations in the EU AI Act concern the regulation of high-risk systems. GPAI models are only implicitly regulated by the systemic risk and serious incident terms. This will have an impact on our further analysis: we focus on high-risk systems compared to GPAI models.

3.3 High-risk AI non-discrimination regulations.

What is a high-risk AI system? Article 6 AIA defines different aspects of high risks. In combination with Annex III, high-risk systems are, for example, intended to be used as safety components in the management and operation of critical digital infrastructure (Annex III(2) AIA) or systems intended to be used for the recruitment or selection of natural persons (Annex III(4)(a) AIA). The reasoning behind the specific regulations in Article 6 AIA is that systems which pose a specific risk to fundamental rights or safety need tighter regulation [Fraser and y Villarino, 2024] than other, less risky systems. Recital 48 states that the adverse impact caused by the AI system on fundamental rights is of particular relevance when classifying an AI system as a high-risk system. It is expected that 5%-15%[8] of all AI systems fall under the category of high-risk systems [Commission, 2021].

We identified Articles 9, 10, and 15 AIA as the main non-discrimination provisions of high-risk systems. These will be discussed in detail in the following paragraphs[9]. Furthermore, Article 11

[7] This classification has been subject to criticism, see e.g., [Bosoer et al., 2023].

[8] For a critical assessment of these numbers see [Almada and Petit, 2023]

[9] Please note that we do not differentiate here between the obligations for providers and those obligations for the deployers of AI systems, since we focus on the interaction of computer science and law.

(together with Annex III) and Article 13 AIA will be discussed due to their relationship to the aforementioned articles.

Article 9 AIA: Non-discrimination regulations within risk management systems. Article 9 of the AI Act covers risk management systems of high-risk systems. According to Article 9(2)(a-d) AIA, risk management systems are designed to identify, evaluate, and mitigate risks of AI systems in a continuous, iterative process throughout the entire lifecycle of a high-risk AI system. All risks after mitigation, called residual risks, need to be judged acceptable [Soler Garrido et al., 2023].

Of particular relevance to non-discrimination law is Article 9(2)(a) AIA, which requires the identification and analysis of known and reasonably foreseeable risks that the high-risk AI system can pose to health, safety, or fundamental rights. Through the link to fundamental rights, indirectly, Article 9 AIA already requires the identification and analysis of non-discrimination issues [Zuiderveen Borgesius et al., 2024]. The initial draft of the AI Act did not include the reference to fundamental rights, and it was added by the EU Council in 2022 [10], highlighting the legislator's intent to include non-discrimination aspects in the AI Act.

Article 10(2)(f) AIA: Main input non-discrimination regulation of high-risk system. The core of the non-discrimination regulation for high-risk systems is Article 10(2)(f) AIA. Article 10(2)(f) AIA requires that training, validation, and testing datasets shall be subject to an "examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations".

The first observation we make is that the input data of an AI system should undergo a bias analysis. Most importantly, the term *bias* is neither defined in the AI Act nor is there a common understanding of its meaning [van Bekkum, 2024]. It seems that the regulators had a more technical definition of bias in mind, focusing on the diversity of training data in different dimensions compared to social, ethical, or structural biases [Hacker et al., 2024b]. This could imply difficulties in determining the regulatory content, also for the later standardisation process.

Through the wording of Article 10(2)(f) AIA, the regulation directly links to traditional non-discrimination law. Many different aspects of discriminatory effects are covered. On the one hand, the link achieves a strong protection objective for the input side of the AI systems. However, even without this link, Union law, which prohibits discrimination, could be applicable (see also next section).

Article 10(2)(f) AIA only targets the input data for machine learning. Under a strict interpretation, the examination may only be mandatory for training, validation, and testing data. In other words, it only covers the input data of an AI system. Although the second half of the sentence reads "especially where data outputs influence inputs for future operations", output data itself is not the main focus.

Furthermore, the wording of Article 10(2)(f) AIA could imply an ex-post view: the developer has to know whether a bias will likely lead to discrimination. However, especially for larger models, it is technically very challenging, if not impossible, to predict the individual output of AI models [Hacker et al., 2024b, Black et al., 2022].

Article 10(2)(g) AIA: Appropriate bias detection, prevention, and mitigation on input data must include factors beyond technical means. Article 10(2)(f) AIA is complemented by Article 10(2)(g) AIA. Article 10(2)(g) AIA demands that "appropriate measures to detect, prevent and mitigate possible biases identified according to point (f)" are considered. Thus, Article 10 AIA not only mandates the examination of biases that lead to discrimination but also the mitigation of these biases. The bias tests must be documented and disclosed according to Article 11 AIA, along with Annexes IV and IXa AIA [Zuiderveen Borgesius et al., 2024]. It has been argued that with this focus on the input side, the AI Act seeks to remedy the root cause of biases [Zuiderveen Borgesius et al., 2024].

Article 10(2)(g) AIA appears to be inspired by the technical literature on fairness metrics for algorithmic outputs. Nevertheless, Article 10(2)(g) AIA targets the input of algorithms. Therefore, classical fairness metrics for algorithmic outputs do not apply directly. Methods for detecting,

[10] See Council of the European Union, "Proposal for a Regulation of the European Parliament and the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach" Doc. 14954/22, 25 November 2022.

preventing, and mitigating input bias remain relatively underexplored in the existing literature, and these methods are not clearly defined in the AI Act [Wachter, 2024, Deck et al., 2024b].

Article 10(2) AIA: Additional obligations without specific fairness regulations. In reviewing Article 10(2) AIA, it is noticeable that the other sections of Article 10(2) AIA also indirectly link to algorithmic fairness considerations. For example, the representativeness of training data is explicitly mentioned in Article 10(2)(d). When interpreting Article 10(2)(f) AIA or Article 10(2)(g) AIA, however, the additional requirements in Article 10(2) AIA do not lead to different results. Since these articles have no direct requirements regarding non-discrimination, we will not analyse them in further detail.

Article 10(5) AIA: A legal basis for the processing of personal data in non-discrimination contexts. Finally, Article 10(5) AIA allows the processing of Article 9 GDPR data (see also Recital 70 AIA). Article 9 GDPR protects special categories of personal data, such as genetic, biometric, or health data (Article 9(1) GDPR). There is tension [Deck et al., 2024a] between the need for debiasing AI algorithms and data protection law, which Article 10(5) AIA aims to address. In order to effectively mitigate biases in AI systems, the processing of personal data (for example, to compute fairness metrics) is important [van Bekkum, 2024]. Notably, this exception only applies in the high-risk regime. Thus, this exception is not applicable to non-high-risk systems. Nevertheless, Article 10(5) is an example of how the AI Act not only imposes burdens on developers and deployers of AI systems but also grants rights to these groups. This aspect is frequently underappreciated in the discourse regarding the AI Act.

Article 11(1) AIA and Annex IV(2)(g) AIA: Technical documentation of bias testing methods. Article 11(1) together with Annex IV(2)(g) AIA requires that the technical documentation of a high-risk system includes "information about the validation and testing data used and their main characteristics; metrics used to measure accuracy, robustness and compliance with other relevant requirements set out in Chapter III, Section 2, as well as potentially discriminatory impacts".

Article 13 AIA: Transparency and provision of information to deployers. Article 13(1) AIA provides that 'High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately". One could contend that the interpretation of a system's output necessarily entails an assessment of potential discrimination. Recital 72 AIA enumerates aspects regarding the information that should be provided, with specific reference to the protection of fundamental rights. However, the risk identified in the recital is presented in general terms and fails to specify any particular methodology or technological tool that must be employed. Moreover, as Recitals serve a non-binding role and do not possess direct normative force in the interpretation of Article 13 AIA, the lack of specificity in Recital 72 AIA does not alter the conclusion. Accordingly, Article 13(1) AIA itself does not impose an obligation to use algorithmic fairness measures or metrics derived from the computer science domain.

Furthermore, even assuming, arguendo, that the use of such algorithmic fairness tools is required during the development or deployment process, Article 13(1) AIA merely obliges providers to document the results. The imposition of mitigation or prevention measures does not fall within the mandatory scope of Article 13(1) AIA.

Article 15 (4) AIA: Output bias regulation in cases of feedback loops. In contrast to Article 10(2)(f) AIA, Article 15(4) AIA mandates that "High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops) and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures." In this case, the output of a system is regulated, in contrast to the input of an AI system in Article 10 AIA. The reasoning behind this is that AI systems should not become echo chambers for biases [Zuiderveen Borgesius et al., 2024]. According to Recital 67 AIA, this is a particular concern when examining historical biases. It is important to note that this output regulation only applies if the systems continue learning after being placed on the market or put into service.

Other articles regarding high-risk systems requirements. While the previously discussed articles provide at least some intuition about the addressed concerns, we categorise the remaining articles of Table 1. Most of these paragraphs only concern fundamental rights impacts, which show a limited impact on non-discrimination requirements. Thus, the discussion here is rather short. Article

13(3)(b)(iii) AIA mandates that information on the foreseeable misuse and the impact on fundamental rights is provided. Article 14(2) AIA mandates that human oversight is needed to minimise risks to fundamental rights. Article 17(1)(i) AIA and Article 26(5) AIA are related to reporting and information on serious incidents. Serious incidents are linked to discrimination through Article 3(49)(c) AIA: "'serious incident' means an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following: [...] the infringement of obligations under Union law intended to protect fundamental rights". Finally, Article 27(1) AIA requires a fundamental rights impact assessment when the high-risk system is used in special systems within bodies governed by public law. Yet again, significant gaps exist between theoretical and methodological elaboration of these impact assessments [Mantelero, 2024]. Article 27(1) AIA only addresses public parties. A fundamental rights impact assessment is not necessary for private parties.

Finally, the impact of fundamental rights on non-discrimination is also monitored by a specific authority established through Article 77 (1) and Article 77 (2) AIA. The effect and impact of this authority framework should be assessed and evaluated once established.

Summary: High-risk systems balancing between specificity and generality. The AI Act aims to be a specific law for AI [Ződi, 2024]. For the ML and Law community, it might be challenging to understand the implications of this specific AI Law. Some Articles mention non-discrimination directly, while others focus on fundamental rights. Only Article 10 AIA and Article 15 AIA explicitly mention non-discrimination and biases, compared to the statement of fundamental rights protection in Article 9 AIA. However, Article 10 AIA only considers the input phase, while Article 15 AIA only addresses the output phase for feedback loops. It is unclear why the AI Act does not consistently regulate direct, indirect, and intersectional discrimination across both the input and output phases.

4 The role of standardisation in the AI Act

According to the New Legislative Framework, all AI systems and models entering the market require a conformity assessment demonstrating that AI systems and models comply with the regulations of the AI Act [Kaminski and Selbst, 2025]. This process will culminate in the establishment of harmonised technical standards [Ebers, 2021].

The details of the conformity assessments are set out in Article 40 ff. AIA[11]. Depending on the specific product, either a self-assessment or a third-party assessment is possible (Article 43 AIA). The AI Act stipulates the requirements for the conformity assessments themselves but does not prescribe the specific content of these assessments [Kaminski and Selbst, 2025].

Through Articles 40 and following, mandates have been issued to the European standardisation organisations. In March 2023, the European Commission issued Mandate M/935 to CEN (European Committee for standardisation) and CENELEC JTC 21 (European Committee for Electrotechnical standardisation) [Kilian et al., 2025]. Thus, the details of non-discrimination procedures —especially concerning methods and the entire risk management framework outlined in Article 17— will be developed through the standardisation process. A similar approach exists for GPAI models, which are governed by the Code of Practice. Experts and stakeholders will collaboratively refine non-discrimination measures for high-risk systems.

However, although standardisation committees have considerable discretion, they must operate within the bounds of legal norms. For example, if the AI Act demands discrimination testing for input data (and excludes output data), standardisation cannot necessarily go beyond this, though it may at times do so. Thus, standardisation in the context of algorithmic fairness can also produce a false sense of safety [Laux et al., 2024].

5 Interplay between traditional EU non-discrimination law and the AI Act for high-risk systems.

With the adoption of the AI Act, questions arise as to how its regime will interact with pre-existing non-discrimination law in the EU. Whereas the AI Act primarily regulates the input side for high-risk systems, much of the machine learning research community's work on fairness concerns algorithmic

[11]For the process of establishing a harmonised standard see [Kaminski and Selbst, 2025].

outputs. As a lex specialis [Craig and De Búrca, 2011], the AI Act prevails over general frameworks to the extent that it covers an issue; where it does not address certain discriminatory outcomes, however, the more general non-discrimination law remains applicable. Since the AI Act does not directly regulate algorithmic outputs for standard high-risk systems, traditional EU non-discrimination law may still apply in these cases.

The relationship between the AI Act and existing fundamental rights protections, including those concerning non-discrimination, requires further clarification and research [Lewis et al., 2025]. This also includes an analysis of when fundamental rights are vertically and horizontally applicable. Furthermore, the fundamental rights in the AI Act need to be balanced [Kusche, 2024], and the material scope is limited due to the enumeration of protected attributes [Xenidis and Senden, 2020]. As pointed out, "gaps in the regulatory framework have left fundamental rights inconsistently protected in the AI Act." [Palmiotto, 2025] Nevertheless, since the AI Act does not protect individuals regarding the output of algorithms, there is room for the applicability of the CFR. It remains an open research question how the protection in the CFR and the AI Act relate [Lewis et al., 2025].

Secondary EU law, such as Directives 2004/113/EC and 2000/43/EC, provides more concrete protections against discrimination, and these instruments have clear horizontal applicability, unlike the frequently debated scope of the CFR [Xenidis and Senden, 2020, Lewis et al., 2025]. For example, Directives 2004/113/EC (Equal Treatment in Goods and Services Directive) and 2000/43/EC (Race and Ethnicity Equality Directive) specifically target non-discrimination regulations. These regulations address specific cases in which discrimination can occur. Due to their specificity, they have a limited material and personal scope as a result of the enumeration of protected attributes [Xenidis and Senden, 2020, Cîrciumaru, 2024]. The development of the relationship between existing regulations and the AI Act, and the stance that national and EU courts will ultimately adopt, are still unresolved. In the past, the ECJ has already issued important rulings on how the non-discrimination provisions in Article 21 CFR must be interpreted in relation to secondary EU law [Xenidis and Senden, 2020]. It is very likely that in the future, the ECJ will continue to issue rulings to clarify the interplay between the EU AI Act, fundamental rights, and other directives and regulations.

The AI Act holds the potential to complement traditional non-discrimination legislation. Traditional non-discrimination law primarily targets the protection of individuals. The EU AI Act is fundamentally aligned with product safety regulation rather than individual rights protection [Almada and Petit, 2023]. Despite the identified challenges associated with this product-safety approach of the AI Act, it nonetheless provides a degree of protection. However, such an emphasis on products can overlook broader regulatory dimensions. In the context of EU data protection, it has been posited that a dual approach involving both individual protection and collaborative governance is necessary [Kaminski, 2018]. A dual approach might also be beneficial for non-discrimination regulations in the context of AI.

6 Summary & Outlook

As [Mayson, 2018] aptly pointed out: "Bias in, bias out." The AI Act encounters significant challenges in regulating such biases. The non-discrimination requirements analysed in the previous section primarily focus on bias in (training) data, and, from our perspective, tend to overlook other sources of bias, such as design choices within the algorithms themselves. Many of the rules emphasise (internal) compliance [Zuiderveen Borgesius et al., 2024] and rely on checklists, rather than offering output-based safeguards for users affected by AI systems. This focus stems from the AI Act's foundations in the New Legislative Framework for product safety law, which, unlike the GDPR, does not incorporate a clear individual rights dimension. Whether self-assessment suffices to protect individuals' fundamental rights and fulfil non-discrimination obligations remains an open question [Bosoer et al., 2023]. This is an area that will require further research, particularly once harmonised standards have been published.

Nevertheless, classical EU non-discrimination law will continue to play an important role in the context of AI. Further research is required to determine how the AI Act's product safety-based regulations can be reconciled with classical, individual-focused non-discrimination law. In the introduction to our work, we raised the question of how fair outcomes from AI algorithms can be ensured. While the AI Act provides some direction, significant challenges persist at the intersection of algorithmic fairness and non-discrimination law. Our paper seeks to bridge the gap between legal and computational disciplines, which, in our view, will benefit from closer interdisciplinary collaboration.

References

Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth. Directly discriminatory algorithms. *The Modern Law Review*, 86(1):144–175, 2023.

Marco Almada and Nicolas Petit. The eu ai act: a medley of product safety and fundamental rights? *Robert Schuman Centre for Advanced Studies Research Paper*, (2023/59), 2023.

Jacy Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Alexander D'Amour, and Chenhao Tan. The impossibility of fair llms. *arXiv preprint arXiv:2406.03198*, 2024.

Shreya Atrey. Illuminating the cjeu's blind spot of intersectional discrimination in parris v trinity college dublin. *Industrial Law Journal*, 47(2):278–296, 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR, 2018.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 602–615, 2024.

Lucía Bosoer, Marta Cantero Gamito, and Ruth Rubio-Marin. Non-discrimination and the ai act. *Law and Digitalization. Arazandi*, 2023.

Stephen Brittain. The relationship between the eu charter of fundamental rights and the european convention on human rights: an originalist analysis. *European Constitutional Law Review*, 11(3): 482–511, 2015.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Olha O Cherednychenko. Fundamental rights and private law: A relationship of subordination or complementarity? *Utrecht Law Review*, 3(2), 2007.

Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. Few-shot fairness: Unveiling llm's potential for fairness-aware classification. *arXiv preprint arXiv:2402.18502*, 2024.

Allessia Chiappetta. Navigating the ai frontier: European parliamentary insights on bias and regulation, preceding the ai act. *Internet Policy Review*, 12(4), 2023.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48, 2024.

Alexandru Cîrciumaru. Discrimination in the age of algorithms—is eu law ready? In *European Yearbook of Constitutional Law 2023: Constitutional Law in the Digital Era*, pages 111–135. Springer, 2024.

EU Commission. Commission notice the 'blue guide' on the implementation of EU product rules 2022 (text with EEA relevance) 2022/c 247/01, 2022. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2022_247_R_0001.

European Commission. Commission staff working document impact assessment accompanying the 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. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021SC0084.

Joseph M Connors, James Michelson, Alex Sudarshan, Sidney Zisook, Kevin Jon Williams, Victoria P Werth, Jon A Wolff, John S Graettinger, and Elliott Peranson. National resident matching program. *New England Journal of Medicine*, 305(9):525–526, 1981.

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846, 2023.

Paul Craig and Gráinne De Búrca. *EU law: text, cases, and materials*. Oxford University Press, USA, 2011.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. Implications of the ai act for non-discrimination law and algorithmic fairness. *arXiv preprint arXiv:2403.20089*, 2024a.

Luca Deck, Astrid Schoemäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. Mapping the potential of explainable artificial intelligence (xai) for fairness along the ai lifecycle. *arXiv preprint arXiv:2404.18736*, 2024b.

Thang Viet Doan, Zichong Wang, Nhat Nguyen Minh Hoang, and Wenbin Zhang. Fairness in large language models in three hours. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5514–5517, 2024.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Martin Ebers. Standardizing ai-the case of the european commission's proposal for an artificial intelligence act. *The Cambridge handbook of artificial intelligence: global perspectives on law and ethics*, 2021.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Matteo Fornasier. The impact of eu fundamental rights on private relationships: direct or indirect effect? *European Review of Private Law*, 23(1), 2015.

Eleni Frantziou. The horizontal effect of the charter of fundamental rights of the eu: rediscovering the reasons for horizontality. *European Law Journal*, 21(5):657–679, 2015.

Henry Fraser and José-Miguel Bello y Villarino. Acceptable risks in europe's proposed ai act: reasonableness and other principles for deciding how much risk management is enough. *European Journal of Risk Regulation*, 15(2):431–446, 2024.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Philipp Hacker. Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under eu law. *Common market law review*, 55(4), 2018.

Philipp Hacker, Johann Cordes, and Janina Rochon. Regulating gatekeeper artificial intelligence and data: Transparency, access and fairness under the digital markets act, the general data protection regulation and beyond. *European Journal of Risk Regulation*, 15(1):49–86, 2024a.

Philipp Hacker, Brent Mittelstadt, Frederik Zuiderveen Borgesius, and Sandra Wachter. Generative discrimination: What happens when generative ai exhibits bias, and what can be done about it. *arXiv preprint arXiv:2407.10329*, 2024b.

David Hadwick and Shimeng Lan. Lessons to be learned from the dutch childcare allowance scandal: a comparative review of algorithmic governance by tax administrations in the netherlands, france and germany. *World tax journal.-Amsterdam*, 13(4):609–645, 2021.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190, 2019.

Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Charles Huff and Joel Cooper. Sex bias in educational software: The effect of designers' stereotypes on the software they design 1. *Journal of Applied Social Psychology*, 17(6):519–532, 1987.

Margot E Kaminski. Binary governance: Lessons from the gdpr's approach to algorithmic accountability. *S. Cal. L. Rev.*, 92:1529, 2018.

Margot E Kaminski and Andrew D Selbst. An american's guide to the eu ai act. *Available at SSRN 5373345*, 2025.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Robert Kilian, Linda Jäck, and Dominik Ebel. European ai standards–technical standardisation and implementation challenges under the eu ai act. *European Journal of Risk Regulation*, pages 1–25, 2025.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

Isabel Kusche. Possible harms of artificial intelligence and the eu ai act: fundamental rights and risk. *Journal of Risk Research*, pages 1–14, 2024.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. *Computer Law & Security Review*, 53:105957, 2024.

Dave Lewis, Marta Lasek-Markey, Delaram Golpayegani, and Harshvardhan J Pandit. Mapping the regulatory learning space for the eu ai act. *arXiv preprint arXiv:2503.05787*, 2025.

Fabian Lütz. The ai act, gender equality and non-discrimination: what role for the ai office? In *ERA Forum*, pages 1–17. Springer, 2024.

Alessandro Mantelero. The fundamental rights impact assessment (fria) in the ai act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review*, 54:106020, 2024.

Sandra G Mayson. Bias in, bias out. *YAle lJ*, 128:2218, 2018.

Kristof Meding and Thilo Hagendorff. Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, 37(1):4, 2024.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8 (1):141–163, 2021.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

Henrik Nolte, Miriam Rateike, and Michele Finck. Robustness and cybersecurity in the eu artificial intelligence act. 2024.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*, 2024.

Francesca Palmiotto. The ai act roller coaster: The evolution of fundamental rights protection in the legislative process and the future of the regulation. *European Journal of Risk Regulation*, pages 1–24, 2025.

Dana Pessach and Erez Shmueli. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer, 2023.

Sacha Prechal. Horizontal direct effect of the charter of fundamental rights of the eu. *Revista de Derecho Comunitario Europeo*, 24(66):407–426, 2020.

Jukka Ruohonen. On algorithmic fairness and the eu regulations. *arXiv e-prints*, pages arXiv–2411, 2024.

Stefan Scheuerer. The fairness principle in competition-related economic law. *GRUR International*, 72(10):919–932, 2023.

Josep Soler Garrido, Delia Fano Yela, Cecilia Panigutti, Henrik Junklewitz, Ronan Hamon, Tatjana Evas, Antoine-Alexandre André, and Salvatore Scalzo. Analysis of the preliminary ai standardisation work plan in support of the ai act, 2023.

Nuno Sousa e Silva. The artificial intelligence act: Critical overview. *Available at SSRN 4937150*, 2024.

Fabienne Ufert. Ai regulation through the lens of fundamental rights: How well does the gdpr address the challenges posed by ai? *European Papers-A Journal on Law and Integration*, 2020(2): 1087–1097, 2020.

Marvin van Bekkum. Using sensitive data to debias ai systems: Article 10 (5) of the eu ai act. *arXiv preprint arXiv:2410.14501*, 2024.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

Sandra Wachter. Limitations and loopholes in the eu ai act and ai liability directives: what this means for the european union, the united states, and beyond. *Yale Journal of Law and Technology*, 26(3), 2024.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567, 2021.

Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. Algorithmic unfairness through the lens of eu non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 805–816, 2023.

Kevin Jon Williams, Victoria P Werth, and Jon A Wolff. An analysis of the resident match, 1981.

Alex Wodi. Artificial intelligence (ai) governance: An overview. *Available at SSRN 4840769*, 2024.

Raphaële Xenidis. When computers say no: towards a legal response to algorithmic discrimination in europe. In *Research Handbook on Law and Technology*, pages 222–234. Edward Elgar Publishing, 2023.

Raphaële Xenidis and Linda Senden. *EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination*. Kluwer Law International, 2020.

Zsolt Ződi. The conflict of the engineering and the legal mindset in the artificial intelligence act. *Available at SSRN 4928715*, 2024.

James Zou and Londa Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.

Frederik Zuiderveen Borgesius, Philipp Hacker, Nina Baranowska, and Alessandro Fabris. Non-discrimination law in europe, a primer. introducing european non-discrimination law to non-lawyers. *Introducing European non-discrimination law to non-lawyers (April 7, 2024)*, 2024.

A Appendix

A.1 Algorithmic fairness: A gentle introduction

Algorithmic fairness has emerged as an established computer science and AI-related field in recent years. Algorithmic fairness focuses on uncovering and rectifying disadvantageous treatment of individuals or groups in machine learning models [Mitchell et al., 2021]. This treatment must be considered within appropriate social, theoretical, and legal contexts [Pessach and Shmueli, 2023] and includes evaluation and auditing methods to test for unfairness.

Origins of unfairness in algorithms. Unfair algorithmic behavior can have different sources. In order to discuss the origins, another — not well-defined — a term from computer science is often used: bias. There exist many different bias definitions. Since the EU AI Act does not define it itself, we use a definition from the EU Commission from 2021. According to this, a bias can be defined as "[...] bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others" [Commission, 2021]. Biases are primarily discussed within the data used for AI models and can be sorted into different categories. Most researchers emphasise the different aspects of data inequalities when discussing algorithmic fairness and sort data biases into different categories, such as the over-representation of specific groups in datasets [Zou and Schiebinger, 2018]. [Hacker, 2018] categorises the biases in data into two groups: biased training data versus unequal ground truth. [Barocas and Selbst, 2016] uses five categories to map biases, while [Mehrabi et al., 2021] uses 21 different categories for biases. It is important to note that the data is only one — albeit important — source of biases. For example, design decisions in the algorithms or structural power asymmetries also lead to algorithmic unfairness [Mehrabi et al., 2021, Gebru et al., 2021, Sousa e Silva, 2024].

Approaches to algorithmic fairness: Quantification. Based on the different understandings of unfairness and for the purpose of uncovering and rectifying unjustified treatment between groups in algorithms, so-called fairness metrics have been proposed [Corbett-Davies et al., 2023, Castelnovo et al., 2022, Verma and Rubin, 2018]. The core idea of fairness metrics is to quantify algorithmic outputs and make them comparable through numerical measurements. Most metrics try to assess whether an output of an ML system is unfair to individuals [Dwork et al., 2012] or groups of people [Barocas et al., 2023, Binns, 2020]. They provide a quantitative measure intended to indicate whether an algorithm demonstrates unjustified unequal treatment. Fairness metrics also include moral norms [Deck et al., 2024a, Hellman, 2020].

The emergence of the field of algorithmic fairness has challenged existing approaches to ML by highlighting the need to contextualize them, decide on, and provide a rationale for the optimization criteria chosen. First, it has been shown that an apparently objective number needs to be contextualized to the application under consideration [Wachter et al., 2021]. Furthermore, some metrics are, from a mathematical viewpoint, incompatible with each other[12], thus these fairness metrics cannot be fulfilled at the same time [Chouldechova, 2017, Kleinberg et al., 2016]. This can lead to an effect of "d-hacking" or "fairness hacking" where users can choose their favorite metric to create the (technical and mathematical) impression that the algorithm is fair [Black et al., 2024, Meding and Hagendorff, 2024]. Fairness metrics for mitigation can play a role at different stages of the development, testing, and deployment of an algorithm. The pre-processing stage, the in-training stage, and the post-processing phases are differentiated [Barocas et al., 2023, Binns, 2018]. In the pre-processing phase, the input data itself is altered to ensure fair processing [Caton and Haas, 2024, Kamiran and Calders, 2012]. For the in-training phase, the optimization process during the training of an ML model is adjusted to include fairness as an optimization goal. Finally, in the post-processing phase, the output of a pre-trained model is adapted. The first and the last approaches make it possible to perform fairness analyses even if one only has black box access to the model [Caton and Haas, 2024].

The era of LLMs raises new questions. LLMs have the advantage — or disadvantage, depending on the viewpoint — that their exact use case is most of the time not predefined, and they are trained on large amounts of various data. This diversity makes them later applicable to different contexts, from providing cookie recipes to coding exercises. However, this diversity also includes variations in unfairness. Algorithmic fairness in relation to LLMs is thus somewhat different from algorithmic fairness in classical AI [Doan et al., 2024, Chu et al., 2024]. Some researchers have applied classical fairness metrics in classification settings to LLMs [Chhikara et al., 2024]. Additionally, it has been shown that LLM-specific issues arise due to the use of massive data and processing [Kotek et al., 2023, Ferrara, 2023, Navigli et al., 2023, Huang et al., 2024] at various steps of algorithmic development. Thus, it was argued that LLMs cannot yield fair outcomes at all [Anthis et al., 2024].

A.2 Table algorithmic fairness related articles

[12]They are also incompatible from a moral point of view [Heidari et al., 2019].

Article	Discrimination	Fundamental Rights	Fair [...]	Bias	Systemic risk	Serious incident	Specific Topic	General Topic
1		x					Subject matter of AI Act	General Provision
2		x					Scope of the AI Act	
3		x			x		Definitions	
6		x					Classification rules for high-risk AI systems	Obligations for High-Risk Systems
7		x					Amendments to Annex III AI Act	
9		x					Risk management system for High-Risk AI Systems	
10	x	x		x			Data and data governance for High-Risk AI Systems	
13		x					Transparency and provision of information to deployers for High-Risk AI Systems	
14		x		x			Human oversight of High-risk Systems	
15				x			Accuracy, robustness and cybersecurity	
17						x	Quality management system	
26						x	Obligations of deployers of high-risk AI systems	
27		x					Fundamental rights impact assessment for high-risk AI systems	
28		x					Notifying authorities	Notifying authorities and notified bodies
36		x					Changes to notifications of the notifying authority	
40		x					Harmonised standards and standardisation deliverables	Standards, conformity assessment, certificates, registration
41		x					Common specifications	
43		x					Conformity assessment	
51					x		Classification of GPAI models as general-purpose AI models with systemic risk	General Purpose AI
52					x		Notification Procedure for GPAI models	
53					x		Obligations for providers of general-purpose AI models	
54					x		Authorised representatives of providers of general-purpose AI models	
55					x	x	Obligations of providers of general-purpose AI models with systemic risk	
56					x		Codes of practice for GPAI Models	
57		x					AI regulatory sandboxes	Measures to support innovation
58		x	x				Detailed arrangements for, and functioning of, AI regulatory sandboxes	
60						x	Testing of high-risk AI systems in real world conditions outside AI regulatory sandboxes	
66		x				x	Tasks of the European Artificial Intelligence Board	Governance
67		x					Advisory forum	
68			x		x		Scientific panel of independent experts	
70		x		x			Designation of national competent authorities and single points of contact	
73						x	Reporting of serious incidents	Post-market monitoring, information sharing and market surveillance
76						x	Supervision of testing in real world conditions by market surveillance authorities	
77	x	x					Powers of authorities protecting fundamental rights	
79		x					Procedure at national level for dealing with AI systems presenting a risk	
82		x					Compliant AI systems which present a risk	
86		x					Right to explanation of individual decision-making	
90					x		Alerts of systemic risks by the scientific panel	
92					x		Power to conduct evaluations	
93					x		Power to request measures	
101					x		Fines for providers of general-purpose AI models	Penalties
112		x					Evaluation and review Article 112 of the AI Act	Final Provisions

Table 1: All articles in the EU AI Act mentioning fairness related terms.

NeurIPS Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We provide an introduction to the AI Act and analyse non-discrimnation regulations for high-risk systems.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We only analyse non-discrimnation regulations for high-risk systems.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: answerYes

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] ,

    Justification: We focus on non-discrimination regulations in the EU AI Act.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.