

Algorithmic Bias in the Context of European Union Anti-Discrimination Directives

Ahmet Bilal Aytekin

Università degli Studi di Genova, Via Balbi 5, Genoa, 16126, Italy

Abstract

The reliance on algorithms for making important decisions instead of humans is widespread, but the expectation for automated decisions to be unbiased is not met. Algorithms have inherited discriminatory behavior from humans and now individuals with protected characteristics face systemic discrimination as a result. To address this pressing issue, the current anti-discrimination laws are studied in this project instead of discussing future regulations. The paper begins by introducing algorithms, machine learning, and automated decision-making and then explains the concept of algorithmic bias. The anti-discrimination laws in the European Union are analyzed to determine the applicability of the legislation (2000/43/EC, 2000/78/EC, 2004/113/EC, 2006/54/EC) in combating algorithmic bias. Although the legislation has limited scope in addressing algorithmic bias, the concept of discrimination, particularly indirect discrimination, can be used to address algorithmic bias in employment, the welfare system, and access to goods and services.

Keywords

algorithmic bias, algorithmic fairness, anti-discrimination law, indirect discrimination


1. Introduction


Algorithms have become an integral part of daily life, even impacting decisions as simple as purchasing shoes online by influencing search results. However, algorithms are not only used in search engines. Because, algorithms are expected to be more efficient, effective, and unbiased, [1] they are used in high-stakes decisions such as predicting criminal recidivism, credit scoring, and job applications. Contrary to common perception, algorithms are not completely bias-free. For example, automated decision-making may rely on biased historical data, which results in discrimination [2]. Consequently, it is imperative to acknowledge that a misapplication in high-stakes decisions can have a more profound impact compared to an erroneous selection of footwear. This is due to the fact that such misapplication can lead to the creation of systemic and far-reaching hazards for individuals who possess protected characteristics, thereby necessitating a comprehensive examination through legal research.


In 2016, ProPublica published an article that reveals the bias in the criminal recidivism prediction algorithm (COMPAS) that is used in the American criminal justice system [3]. The article drew attention which resulted in the passing of a bill by New York City to assign a task force to examine algorithmic bias in automated-decision making programs used by

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

✉ abilalaytekin@gmail.com (A. B. Aytekin)

ORCID  0000-0002-7911-2157 (A. B. Aytekin)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

government agencies[4]. There have also been significant advancements and progress made toward regulating artificial intelligence (AI) in Europe. The European Union (EU) has taken a leading role in regulating the use of AI with a focus on risk management. The EU AI Act, which is currently in parliament and expected to be enacted by 2024, aims to set global standards for AI regulation. This legislation is part of a broader effort to address the implications of AI adoption and includes provisions in the Digital Markets Act (DMA) and Digital Services Act (DSA) calling for algorithmic transparency through independent audits. Although the EU's regulation of AI is a positive step, it may be viewed as limited in scope. Nevertheless, the combined effect of the EU AI Act, DMA, and DSA is expected to have a significant impact on the regulation of algorithms and AI in business and organizational practices.[5]

On the other hand, in the European Union (EU), the General Data Protection Regulation (GDPR) is one of the applicable legislation addressing automated decision-making. Article 22 of the GDPR entitles the right not to be subject to a decision based solely on automated processing and addresses discrimination, which derives from the use of sensitive data, in profiling. Considering especially Recital 71 and Article 22(4), the GDPR gives the impression of being a reliable safeguard against automated decision-making [6]. While the GDPR is considered a significant step towards protecting individual privacy rights and data security in EU, there are some who have expressed skepticism regarding its efficacy in addressing issues of algorithmic bias and discrimination. The prospect of implementing a legally binding right to explanation as a safeguard against automated decision-making within the framework of the GDPR faces numerous significant challenges. To fall under this framework, an automated decision-making process must rely "solely on automated processing" and produce "legal effects" or similarly significant consequences [7]. The ambiguous and limited scope of Article 22's 'right not to be subject to automated decision-making,' which serves as the foundation for the claimed 'right to explanation,' casts doubt on the actual protection provided to data subjects. These reservations suggest that the GDPR may be lacking in precise language and explicit, well-defined rights and safeguards against automated decision-making, potentially leading to a less effective regulation [6]. Therefore some scholars have argued that the interpretation of the GDPR suggests that its approach to addressing discrimination caused by algorithmic bias is either ineffective or unfeasible[8, 9]. Furthermore, the GDPR's accountability mechanisms have also been criticized for their inadequacy in ensuring transparent and accountable automated decision-making[6].

However, it is essential to note that evaluating the effectiveness of legal regulations is a complex matter, and it is beyond the scope of this discussion to delve into this topic in detail. This point is made to highlight the fact that it is not always straightforward to rely on the title or apparent purpose of a legal rule to determine its efficacy in practice. Consequently, given the ongoing need for a legal framework that effectively addresses the multifaceted challenges posed by algorithmic bias and discrimination, it may be advantageous to investigate alternative legislation. As a distinguished international entity, the European Union (EU) has demonstrated unwavering commitment to confronting and mitigating various manifestations of prejudice, recognizing the inherent importance of cultivating equitable treatment and social justice. As a result, the EU's well-established frameworks and anti-discrimination legislation could be a valuable asset in addressing the complex issue of algorithmic bias, potentially providing key insights and strategies to mitigate its widespread consequences. A thorough examination of these frameworks may highlight the potential relevance of anti-discrimination laws in

establishing a more appropriate legal foundation for addressing the pressing concerns about algorithmic bias and discrimination.

Additionally, I will not provide a detailed explanation of algorithmic bias in this introduction. Instead, this topic will be the focus of the next chapter, which will delve into the definition, causes, and impact of algorithmic bias in greater depth.

The objective of this paper is to investigate the potential role of EU anti-discrimination law in addressing discrimination caused by algorithmic bias. To achieve this aim, the paper will first establish a clear understanding of the relevant terminology and concepts related to algorithmic bias. Subsequently, it will examine the various forms of discrimination and the scope of EU anti-discrimination legislation, with the intent of assessing its applicability to instances of algorithmic bias. Through this investigation, the paper aims to contribute to the ongoing discourse on the legal and regulatory obstacles posed by algorithmic bias

2. Background

2.1. Algorithm, machine learning, automated decision-making

It is impractical to provide comprehensive definitions or elucidations for all of the technical terminology relevant to this domain of in the limited scope of a concise article. Nonetheless, it is critical to familiarize the reader with key technical terms and concepts, particularly those that will be referenced in subsequent sections, in order to facilitate a more in-depth understanding of the subject matter. By briefly explaining these key terms, we hope to elucidate the complex ideas underlying the topic, fostering a better understanding of the interconnected concepts and a more robust engagement with academic discourse. Briefly, an algorithm is thus a sequence of computational steps that transform the input into the output[10]. On the other hand, a computer algorithm is a special type algorithm which is “a set of steps to accomplish a task that is described precisely enough that a computer can run it”[10]. A computer algorithm consists of two components which are: logic and control. Firstly, the logic is “the problem domain-specific component and specifies the abstract formulation and expression of a solution”[11]. Secondly, the control is the component which is “the problem-solving strategy and the instructions for processing the logic under different scenarios”[11]. A computer algorithm should produce a correct solution to the problem and while producing, it must be efficient. If a computer algorithm presents incorrect solutions or provides correct solutions but in an inefficient way, it means that the algorithm has little or no value[10]. The improvement of the efficiency of a computer algorithm lies within the refinement of the two components[11]. For the purposes of this discussion, a brief explanation of algorithms is sufficient, and it is important to note that the term "algorithm" will now be used specifically to refer to a "computer algorithm."

The other related concept is “machine learning” which can be described as “programming computers to optimize a performance criterion using example data or experience”[12]. Machine learning uses the theory of statistics and computer science. The primary purpose of machine learning is to make an inference from a sample. Hence, the theory of statistics is used to build mathematical models. Additionally, computer science is mainly used for two purposes. First, because, there is a significant amount of data to store and process, efficiency is required and computer science gets involved to solve the optimization problem. Second, a model's

representation and algorithmic solution for inference requires to be efficient also, after the model is completed[12]. Machine learning is divided into three categories (Supervised, semi-supervised and unsupervised) which are grouped according to the nature of the data labelling. If all data is labelled and machine learning is used for the estimation of an unknown mapping, it is called supervised learning. In unsupervised learning there is no labelled data, only input samples are given. In semi-supervised, the data is partially labelled and it is used for inferring the unlabelled part[13].

Automated decision-making refers to decision-making based on statistical models or decision rules without explicit human intervention[14, 15]. Prioritisation, classification, association and filtering are the four main types of decisions that can be made by an automated system[16]. Firstly, the purpose of prioritisation is to emphasize or underline determined things at the expense of others[16]. Secondly, classification is used for “categorising a particular entity as a constituent of a given class by looking at any number of that entity’s features”[16]. Thirdly, association decisions serve to designate relationships between entities[16]. Lastly, filtering is employed for “including or excluding information according to various rules or criteria”[16]. Taking everything into account, this section only provides a brief overview of these concepts, but it should be adequate to move forward with the topic of algorithmic bias.

Finally, while the explanations provided are by no means exhaustive, they should suffice to facilitate a more profound understanding of the subject as we delve deeper into the intricate nuances of algorithmic bias and its implications. It is our intention that this foundational knowledge will encourage meaningful engagement with the subsequent academic discourse and contribute to a more comprehensive comprehension of the broader thematic landscape.

2.2. What is algorithmic bias?

This chapter delves into the concept of algorithmic bias, which has received a lot of attention in recent years. However, debates over the term have frequently been complicated by differing interpretations and usages. There are numerous definitions for algorithmic bias, which merits further investigation. Automated decision systems, particularly those that use machine learning classification models, are inherently designed to discriminate—that is, to detect differences [17, 18]. This does not, however, imply that all such systems are inherently biased. According to Friedman and Nissenbaum, to manifest as an algorithmic bias in automated decision systems, discrimination must be both systematic and unfairly oriented towards certain groups or individuals at the expense of others [19]. An automated system discriminates unfairly "if it denies an opportunity or a good or assigns an undesirable outcome to an individual or group of individuals on unreasonable or inappropriate grounds" [19]. It is important to note that unfair discrimination caused by random system errors does not constitute algorithmic bias. Unfair discrimination must occur systematically in order to be considered algorithmic bias [19]. In contrast, unless systematic discrimination results in inequitable outcomes, it cannot be classified as algorithmic bias [19]. For instance, the EU anti-discrimination law allows setting quotas on the side of under-represented groups. Accordingly, systematic preference of women in a workplace to reverse discrimination does not have an unfair outcome, because it promotes substantive equality[20]. Therefore, positive discrimination in an algorithm does not constitute algorithmic bias.

Examining alternative conceptualizations of algorithmic bias reveals that some scholars prefer the term "unfair" rather than "biased." This choice is motivated by the desire to reserve the term "biased" for its original statistical connotations, while using "unfair" to encompass the phenomenon's social and moral dimensions [21]. Mehrabi and others asserts that fairness in the domain of automated decision-making is defined by the absence of any partiality or prejudice that may arise from innate or acquired traits [22, 21]. Barocas and others on the other hand, avoid the term "bias" in favor of phrases like "demographic disparity" and "discrimination" to explain the negative effects of specific computational models. They preserve the term "bias" for its traditional statistical meaning of systematic error by taking this approach[23, 21].

To avoid delving deeper into the myriad definitions of algorithmic bias, I will concentrate on the Friedman and Nissenbaum conceptualization. This method allows us to maintain a coherent and focused discussion on the topic at hand while drawing on the insights provided by these prominent field scholars. Algorithmic bias can be broadly classified into three distinct categories: pre-existing bias, technical bias, and emergent bias. First, pre-existing bias originates from societal practices and attitudes that are external to and independent of the automated system's coding process [19]. Historical biases present in input data offer an apt illustration of pre-existing bias. For example, the predictive policing algorithm PredPol disproportionately targets African-Americans due to its reliance on police records, which embody biases against this demographic [24]. Second, technical bias emerges as a consequence of factors or limitations encountered during the design and development process of an automated system. Such biases may be attributed to constraints in computer tools or the use of context-specific algorithms [19]. The efficacy of a system is contingent upon both hardware and efficient algorithms [10]. In the pursuit of optimizing performance and efficiency, not all relevant information may be considered, potentially resulting in unmeasured confounding factors and biased outcomes [25]. Finally, emergent bias pertains to biases that materialize after algorithms have been deployed and are subject to changes in their contextual environment [19]. A notable example is Microsoft's "social chatbot," Tay, which was designed to engage in amicable conversations with users. However, Tay began to mimic the malicious speech patterns of certain users, ultimately transforming into a racist chatbot. This incident compelled Microsoft to issue an apology and discontinue Tay's operation, as it had propagated offensive statements, such as comparing former President Obama to monkeys and denying the occurrence of the Holocaust [26]. In summary, these three categories of algorithmic bias underscore the myriad ways in which biases can infiltrate and influence automated systems, necessitating a comprehensive understanding of their origins and potential ramifications.

Discrimination in automated systems is primarily caused by factors such as biased data, underrepresentation of specific groups in data, statistical models, or decision-making rules. As previously stated, automated systems rely on datasets to identify patterns that guide decision-making. These datasets used during the design stage are commonly referred to as training data in the field of computer science [27]. Training data may inadvertently include societal biases, which then manifest in the system's results [2]. This is an excellent example of pre-existing bias. Training data may not necessarily contain historical biases, but it may lack adequate representation of specific demographics, such as race, gender, or age. This underrepresentation can lead to algorithmic bias in automated systems, particularly in classification tasks [28]. Inadequate representation has a direct impact on data quality, which is defined as the degree to

which data meets the requirements for a specific purpose [29]. High-quality decision-making is dependent on data quality, as it is difficult to make well-informed decisions in the absence of reliable data. Furthermore, statistical models or decision rules underpin automated decision-making algorithms. In some cases, the use of specific models or rules may also contribute to the appearance of algorithmic bias [30]. This demonstrates the complex interplay of factors that can lead to biased results in automated systems.

It is crucial to acknowledge that the origins of algorithmic bias extend beyond the factors delineated in this section. Indeed, there may be an array of additional elements that contribute to the emergence of such biases. This observation underscores the multifaceted nature of algorithmic bias and the necessity for continued exploration and understanding of its potential sources within the context of automated systems.

3. The legal framework for prevention of discrimination in the european union law

So far, I have attempted to clarify the technical aspects of algorithmic bias. In this section, I will analyse the European Union's anti-discrimination law, specifically the Discrimination Directives, to determine whether it provides a legal foundation for addressing algorithmic bias. The legal order of the European Union is divided into three groups: primary legislation, secondary legislation, and supplementary law. The principle of equality and prohibition of discrimination found its place in all types of sources of EU law. As the primary source, the Treaty on European Union Article 2 states that equality is one of the founding values of the European Union. Moreover, in Article 3(3) of TEU, combating discrimination and promoting equality are given as a goal of the EU, similarly in Article 10 of the Treaty on the Functioning of the European Union emphasized that "... the Union shall aim to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.". Furthermore, another significant source for the anti-discrimination law is the Charter of Fundamental Rights of the European Union which "has the same legal value in EU law as the founding Treaties"[20]. According to Article 21 of the Charter: "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 shall be prohibited". Article 19(1) of the TFEU authorises all types of legislation or other instruments for the prohibition of discrimination[20]. As the secondary sources of law;

- Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Racial Equality Directive),
- Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation (Equality Framework Directive 2000),
- Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Equal Treatment in Goods and Services Directive) and
- Directive 2006/54/EC of the European Parliament and of the Council on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (Equal Treatment Directive 2006)

are enacted in the EU. The EU anti-discrimination directives aim to prevent discrimination on different grounds. In the next section, the protected grounds which are covered by the Directives are demonstrated.

3.1. The protected grounds

The protected grounds are more comprehensive in the primary sources of law of the EU, comparing the anti-discrimination directives. According to Article 19 of TFEU; sex, racial or ethnic origin, religion or belief, disability, age, and sexual orientation can be considered protected grounds. Additionally, Article 18 of the TFEU prohibits any discrimination on the grounds of nationality. Moreover, the CFREU provides a wider range of protection. Besides the above-mentioned grounds, Article 21 of CFREU also includes, social origin, genetic features, language, political or any other opinion, membership of a national minority, property, and birth. Additionally, an anti-discrimination directive provides a prohibition on the protected ground, if a protected ground is covered by the Directive. The Race Equality Directive Articles 1 and 2 prohibits discrimination only on the grounds of “race or ethnic origin”. According to Article 1 of the Equality Framework Directive 2000, religion or belief, disability, age, and sexual orientation are the protected grounds. Sex is the common protected ground in the Equal Treatment in Goods and Services Directive and Equal Treatment Directive 2006. The anti-discrimination directives categorize discrimination into two types: direct and indirect discrimination. The next section will delve into these two categories of discrimination.

3.2. Categories of discrimination

3.2.1. Direct discrimination

Direct discrimination means “unfavourable or less favourable treatment on the ground of a protected characteristic (such as race, sex, religion) or, sometimes, a combination of such characteristics” [31]. In the EU anti-discrimination directives, direct discrimination is similarly defined. For instance, in Article 2 of Equal Treatment Directive 2006, direct discrimination entails a circumstance “where one person is treated less favourably on the grounds of sex than another is, has been or would be treated in a comparable situation”. It can be argued that the prohibition of direct discrimination aims to preserve the principle of formal equality [20]. The definition of direct discrimination indicates three elements. The first, direct discrimination requires a “less favourable treatment” which could be doing something or the failure to do something [31]. The “less favourable treatment” would be receiving a lower wage, being rejected for a job, a refusal for a social benefit, etc. Second, there is a need for a comparator, to evaluate the “treatment”. “To treat someone as different means to accord them a treatment that is different from the treatment of someone else; to describe someone as “the same” implies “the same as” someone else” [32]. Accordingly, the “difference” in treatment is only meaningful as its comparison to others [32]. Therefore, the claimant must demonstrate the less favourable treatment in comparison with an individual or group who is in a similar situation [31]. Third, the less favourable treatment must base on a protected characteristic [31] such as race, gender or age.

Another issue that needs to be addressed is whether direct discrimination can be justified under EU anti-discrimination law or not. According to Evelyn Ellis and Philippa Watson:

“... There are, therefore, broadly speaking, two elements of the tort of discrimination, whichever form it take: adverse treatment (harm) and the grounding of that treatment in a prohibited classification (causation)... It has also been seen that objective justification reflects the element of causation where the discrimination is indirect: if the adverse consequences to one group can be shown to be attributable to an acceptable and discrimination-neutral factor, then there is no discrimination. The cause of the adverse impact is something other than discrimination. When one is dealing with direct discrimination, however, once adverse treatment and causation have been proved, this is the end of the matter; there can logically be no room for any further arguments about the roots of the adverse treatment. Justification is, therefore, not an applicable notion.”[20]

However, in some circumstances, it is possible to give a justification for direct discrimination where the EU anti-discrimination directives allow. For example, Article 4(5) of the Goods and Services Directive states that: “This Directive shall not preclude differences in treatment, if the provision of the goods and services exclusively or primarily to members of one sex is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary”. Overall, it is not wrong to say that the general defence of justification is exceptional for direct discrimination cases.

3.2.2. Indirect discrimination

Equal treatment does not necessarily produce equal outcomes for different groups[33]. Accordingly, this idea gave rise to the theory of indirect discrimination (disparate impact) which is developed by the United States Supreme Court in the case of *Griggs vs Duke Power*. The prohibition of indirect discrimination is invoked where “an unjustified adverse impact is produced for a protected class of persons by an apparently class-neutral action”[20] to sustain substantive equality, particularly the principle of equal opportunity, rather than formal equality[20]. Thus, “the law of indirect discrimination tackles problems of social integration and social inclusion by ensuring that disadvantaged groups in society do not encounter nearly insuperable obstacles in becoming integrated through education, work, and participation in social life”[34]. The prohibition of indirect discrimination is accepted in the EU anti-discrimination law, and it is defined in the anti-discrimination directives. For example, in Article 2 of the Equal Treatment Directive 2006. Because this directive involves gender equality in matters of employment and occupation, race is not mentioned in the definition. However, it is defined in the Race Equality Directive Art. 2(b) that mentions “race” and “ethnic origin”.

It is specified within the previously cited Articles that indirect discrimination encompasses four distinct elements. The first element is that there must be an apparently neutral provision, criterion or practice that is applied to anybody, in other words, “there must be equal treatment”[33]. The second, an apparently neutral provision, criterion or practice which has a negatively disproportionate impact on a protected group, is necessary[31]. Basically, indirect discrimination concerns “with impact rather than treatment”[33] which applies equally to all subjects. The third element is the necessity of a comparator. For indirect discrimination, comparisons involve groups rather than individuals[31]. For constructing an indirect discrimination

case, it is essential to determine a group of people as the comparator for proving that the claimant who belongs to a protected group received significantly less advantageous treatment in its effect[20]. The use of statistics may necessarily be determining the questions related to the second and the third element[33]. Choosing the right comparator is as important as showing statistical evidence. For example, when a woman claims that she is discriminated on the grounds of her sex, then choosing men as the comparator allows her to present strong arguments proving her exclusion[32]. However, in some cases it is challenging to find an appropriate comparator. For instance, pregnancy is only attributed to women, therefore, pregnancy is not a comparable situation with men[32]. As a result, in the EU anti-discrimination law, it is not necessary to find a comparator for proving discrimination due to pregnancy[20]. The fourth element is that indirect discrimination is always justifiable,[33, 31] if the provision, criterion or practice, is a necessary and proportionate tool for the legitimate purpose[31]. Ellis and Watson explain more explicitly:

“The respondent must show to the satisfaction of the national court that there is a genuine need on behalf of the enterprise for the discriminatory factor, that the means chosen are suitable for attaining the objective, and, most strictly of all, that the means chosen are ‘necessary’ to attain the objective; it follows that, if reasonable alternative means are available to the respondent to attain the objective, the behaviour will breach the non-discrimination principle.” [20]

After all, it can be contended that while there is no general defence for direct discrimination, it is possible to justify indirect discrimination.

3.3. The scope of the EU anti-discrimination directives

First, it is essential to answer the question of who is protected under the anti-discrimination directives. As a rule, the anti-discrimination directives are “intended to apply to all persons within the EU, irrespective of their nationality”[20]. However, the Race Equality Directive and the Equality Framework Directive exclude some aspects of the protection of nationals of third countries. Article 3(2) of Race Equality Directive and Equality Framework Directive:

“This Directive does not cover the difference of treatment based on nationality and is without prejudice to provisions and conditions relating to the entry into and residence of third-country nationals and stateless persons on the territory of Member States, and to any treatment which arises from the legal status of the third-country nationals and stateless persons concerned.”

Also, the Recital 12 of the preamble of the Race Equality Directive excludes the application of the directive in matters of access to employment and occupation for third-country nationals. On the contrary, in Equal Treatment in Goods and Services Directive and Equal Treatment Directive 2006, there is no exclusion of third-country nationals. The scope of the Racial Equality Directive involves:

- a) conditions for access to employment, to self-employment, and to occupation, including selection criteria and recruitment conditions, whatever the branch of activity and at all levels of the professional hierarchy, including promotion

- b) access to all types and to all levels of vocational guidance, vocational training, advanced vocational training, and retraining, including practical work experience
- c) employment and working conditions, including dismissals and pay
- d) membership of and involvement in an organisation of workers or employers, or any organisation whose members carry on a particular profession, including the benefits provided for by such organisations
- e) social protection, including social security and healthcare
- f) social advantages
- g) education
- h) access to and supply of goods and services which are available to the public, including housing

The Equality Framework Directive only covers (a), (b), (c), and (d) as the scope of the directive. On the other hand, according to Article 3 Equal Treatment in Goods and Services Directive, the scope of the prohibition on discrimination includes access to the supply of goods and services, excluding the content of media and advertising; education; matters of employment and occupation. Lastly, the Equal Treatment Directive 2006 applies to: “access to employment, including promotion, and to vocational training; working conditions, including pay; occupational social security schemes”. To sum up, it can be said that the Directives offer protection in three main areas which are, employment, the welfare system, and access to the supply of goods and services.

4. Legal analysis of discrimination caused by algorithmic bias

This paper has thus far attempted to explain the concept of algorithmic bias as well as the legal framework of EU anti-discrimination law. In this section, I will look at algorithmic bias in the context of EU anti-discrimination legislation. First, I will investigate the classification of discrimination caused by algorithmic bias, debating whether it is direct or indirect discrimination. Following that, I will discuss whether discrimination caused by algorithmic bias can be justified. Finally, I’ll share my thoughts and observations on the subject.

As previously stated, there must be a discernible difference in treatment, whether unequal or less favorable, directed towards an individual or a group based on a protected characteristic for an act to be classified as direct discrimination. Indirect discrimination, on the other hand, requires the presence of an ostensibly neutral provision, criterion, or practice that is applied uniformly to all individuals. It is widely acknowledged that automated systems have an ostensibly objective personality, which contributes to their perceived neutrality [35]. Furthermore, automated systems are generally designed to be applicable to all individuals within their designated scope of application. For instance, a credit scoring algorithm is employed across the board for all loan applicants by a bank or credit institution. Taking these characteristics of automated systems into account, it appears more plausible to categorize algorithmic bias as a manifestation of indirect discrimination. However, according to Collins and Khaitan, it is not straightforward to differentiate direct and indirect discrimination, since the distinction relies on not one but multiple differences[34]. One of the differences is that indirect discrimination is always concerned with groups. Contrarily, direct discrimination is commonly affecting individuals (although not

necessarily)[34]. Another way to distinguish direct discrimination from indirect discrimination is by looking at the exclusionary effect of discriminative practice or rule. Collins and Khaitan suggest that:

“Direct discrimination is usually defined as the adoption of a ground for decision that will exclude 100 percent of the protected group, but none of its cognate and comparative group. It follows that indirect discrimination applies where the exclusionary effect of the practice or rule is less than 100 percent but is disproportionate in comparison to cognate groups”[34]

Because of the fact that, in the EU law, proving intention or motive is not necessary to elucidate direct or indirect discrimination[20]. Using the effect of a discriminative act as a criterion makes the distinction more clear [34]. Another difference is the possibility of justifying indirect discrimination. Unlike indirect discrimination, the general defense of the proportionality test or “business necessity” test is not applicable to direct discrimination claims[34].

It is critical to understand that automated decision-making systems are not designed to target specific individuals; rather, they are implemented to evaluate larger populations, such as potential employees or loan applicants. As a result, it can be argued that a flawed automated system always affects a substantial number of people. Another distinguishing feature is the varying exclusionary effects of algorithmic bias across different systems. In the absence of explicit intent to marginalize one group in favor of another, the exclusionary impact on protected groups is unlikely to reach total exclusion. This variation in exclusionary effects emphasizes the nuanced nature of algorithmic bias and its differing implications in various decision-making contexts. For example, ProPublica’s article on machine bias demonstrates that Northpointe’s criminal recidivism assessment tool correctly predicts 61 percent. However, African-Americans are more likely to be labeled as a higher risk, although they do not re-offend, comparing whites[3]. In most cases, the general defence of justification can be used by the stakeholders, since the very first reason for employing an automated system is to have more effective, efficient, consistent, and unbiased results[36]. Taking these distinctions into account, it appears increasingly reasonable to classify discrimination resulting from algorithmic bias as a type of indirect discrimination. This definition recognizes the subtle and often unintentional nature of algorithmic bias, while also emphasizing its potential to produce disparate outcomes for protected groups in various decision-making contexts. We can better understand and address the underlying mechanisms that contribute to algorithmic bias-induced discrimination by framing it as indirect discrimination, ultimately fostering a more just and equitable application of automated decision-making systems.

As a result of classifying algorithmic bias as a form of indirect discrimination, the common defense of justification becomes applicable in such cases. The test of proportionality is the dominant standard for assessing justification in the context of EU anti-discrimination law[34]. The proportionality test is an important criterion for determining the legitimacy of measures that may unintentionally result in disparate outcomes for protected groups. The fundamental goal of indirect discrimination law, as explained in the preceding section, is to advance substantive equality by ensuring that the pursuit of fairness is not jeopardized by the unintended consequences of algorithmic biases. On this account, Collins states that:

“... neutral rules and practices that disproportionately exclude members of disadvantaged groups are not in themselves evidence of a moral wrong, but to the extent that those rules and practices operate to obstruct social mobility and exacerbate social exclusion, they are likely to be regarded as rules and practices that should be discouraged and if possible, at a reasonable cost, eliminated”.

As a result, the proportionality test should be viewed as a comprehensive assessment that weighs the costs and benefits of removing systemic barriers to social inclusion. This analytical approach allows for a more nuanced understanding of the trade-offs involved in overcoming such impediments, ensuring that the pursuit of substantive equality is both effective and mindful of potential unintended consequences. Legal frameworks can strike a delicate balance between promoting social inclusion and preserving the functional integrity of automated decision-making systems by using a cost-benefit analysis within the context of the proportionality test [34]. To address algorithmic bias, two competing principles must be optimized, striking a delicate balance between the rights of affected individuals and the legitimate interests of institutions deploying these algorithms. A loan applicant, for example, has the fundamental right to be free of discrimination in the context of a biased credit scoring algorithm. In contrast, the bank or credit institution retains the right to exercise professional judgment in determining creditworthiness. Legal frameworks must carefully consider the broader implications of their decisions in order to promote substantive equality while preserving the operational viability of institutions reliant on automated decision-making processes.

Robert Alexy emerges as a leading authority in the field of constitutional rights theory and the test of proportionality. His second book, *A Theory of Constitutional Rights*, published by Nomos in German in 1985, quickly rose to prominence as the most influential postwar German book on constitutional rights theory. Considering the profound influence of German constitutional rights doctrine on the examination of European fundamental rights, particularly in the context of the Charter of Fundamental Rights of the European Union, Alexy's theoretical approach has piqued the interest and scholarly attention of scholars across Europe. [37] According to Alexy's theoretical framework, optimization within the realm of factual possibilities entails circumventing avoidable costs. However, when principles clash, costs become unavoidable, necessitating the implementation of a balancing process [38]. In my view, adopting an economically oriented approach to this subject can yield more equitable outcomes, taking into account the rights of all involved parties. I consider Robert Alexy's theory (the principle of proportionality) to be the most fitting paradigm for the test of proportionality, as it strives for Pareto optimality in reconciling competing principles. This approach ensures that any adjustments made to one principle result in a net benefit without causing undue harm to the other, thereby fostering a harmonious balance between conflicting imperatives. According to Alexy, the principle of proportionality is composed of three sub-principles which are: the principles of suitability, of necessity, and of proportionality in the narrower sense [38]. Firstly, the principle of suitability, “precludes the adoption of means that obstruct the realization of at least one principle without promoting any principle or goal for which it has been adopted” [38]. The second sub-principle, the principle of necessity, “requires that of two means promoting one of the competing principles that are, broadly speaking, equally suitable, the one that interferes less intensively with the other competing principle has to be chosen. If there exist a less intensively interfering

and equally suitable means, one position can be improved at no costs to the other”[38]. The third sub-principle, the principle of proportionality in the narrower sense, is about balancing the principles. Conformably, Collins claims that the function of justification is to balance the rights of the parties[34]. This sub-principle is similar to the rule of "Law of Balancing" which states: "The greater the degree of non-satisfaction of, or detriment to, one principle, the greater must be the importance of satisfying the other". The "Law of Balancing" necessitates further explanation in order to conduct a thorough and comprehensive analysis, which culminates in the development of the Weight Formula¹ [38]. As a result, in order for a biased system to pass the proportionality test, it must adhere to all three sub-principles. This adherence ensures that competing concerns are evaluated in a systematic and nuanced manner, allowing for a more equitable resolution of conflicts arising from algorithmic bias and indirect discrimination within decision-making processes.

The primary motivation for using algorithms is to accelerate up and improve data processing capabilities. As a result, it is possible for a defendant, particularly in the private sector, to build their defense on the premise of "business necessity." Using a biased credit scoring algorithm as an example, a bank or credit institution may argue that using an algorithmic system promotes efficiency, resulting in improved service quality and increased customer satisfaction. As a result, the bank may argue that the measure (i.e., the algorithm) advances its freedom to practice its trade. It is possible to argue that the principle of suitability is met in cases of algorithmic bias. It is worth noting that instances of noncompliance with the principle of suitability are uncommon. [38]. On the other hand, analysing the proportionality in the narrower sense for the cases of algorithmic bias is more controversial than the principle of suitability. Applying "Law of Balancing" consist of three stages:

"The first stage involves establishing the degree of non-satisfaction of, or detriment to, a first principle. This is followed by a second stage in which the importance of satisfying the competing principle is established. Finally, in the third stage, it is established whether the importance of satisfying the latter principle justifies the detriment to or non-satisfaction of the former" [39]

The facts of the relevant case are crucial to apply the "Law of Balancing" hypothetical application of the rule of "Law of Balancing" does not produce absolute results, because, balancing is "a conditional relation of precedence between the principles in the light of the circumstances of the case"[38]. As a result, it is impossible to draw a definitive conclusion regarding the test of proportionality in the narrower sense without access to specific details of an actual case, such as the number of customers served by the bank, the extent of disproportionate treatment, or the benefits accrued by the bank. In some cases, a bank may meet the proportionality criterion in the narrower sense, while in others, it may not. Importantly, I argue that using a biased algorithmic system consistently violates the principle of necessity. The root cause of discrimination is not the use of an algorithmic system; rather, it is the use of a biased one. As previously stated, potential sources of discrimination within a system include the use of biased training data, underrepresentation in the training data, biased statistical models, and the

¹See R. Alexy, Constitutional rights and proportionality, *Revus. Journal for Constitutional Theory and Philosophy of Law* (2014)

decision rule used. Importantly, these negative characteristics can be detected [40] and then rectified or eliminated [41, 29, 42, 43, 44]. Assuming that using a biased algorithm is not an absolute necessity, deploying a biased system consistently fails to meet the principle of necessity, ultimately failing the proportionality test. As a result, discrimination resulting from algorithmic bias should be regarded as indirect discrimination under EU anti-discrimination law.

In general, redress for indirect discrimination that fails the proportionality test entails the repeal or modification of the offending provision, criterion, or practice in a way that mitigates its disproportionate impact on the protected group[26]. In contrast, proving direct discrimination usually results in compensation being awarded[29]. It is worth noting that, in the absence of demonstrable discriminatory intent on the part of the defendant, the allocation of damages in cases of indirect discrimination is relatively uncommon[26]. This distinction emphasizes the significance of carefully assessing the nature of discrimination and the appropriate remedies in the context of legal proceedings.

However, it is important to note that the preceding explanation focuses on one specific aspect of the difficulties associated with the implementation of anti-discrimination directives, namely the justification of indirect discrimination. Other issues, such as group identification[45], may complicate the classification of automated decision-making as indirect discrimination and necessitate further investigation. A thorough understanding of the nuances of discrimination in the context of algorithmic systems necessitates a multifaceted approach that takes into account all of the facets and potential pitfalls that may arise during the implementation and evaluation of anti-discrimination efforts.

In summary, it can be inferred that discriminatory outcomes resulting from algorithmic bias fall under the category of indirect discrimination, thereby invoking the protective provisions of the EU anti-discrimination directives within its limited scope. However, given the ubiquity of automated decision-making, its reach and impact transcend the confines of the Directives. Consequently, it is apparent that the existing framework falls short of providing comprehensive safeguards against algorithmic bias in diverse sectors. The scope of the Directives only covers three main areas, employment, the welfare system, and access to the supply of goods and services. In addition, it should be noted that the scope of the Directives is limited, as they do not apply to certain areas, such as education under the Equal Treatment in Goods and Services Directive. This implies that an algorithm in a school admission process that has a disproportionate effect on a specific gender would not be deemed illegal within the ambit of the Directive. Similarly, a discriminatory algorithm used by LinkedIn or other third-party websites, which is not an employer, would not be contestable under the EU Directives. Such limitations in the scope of the Directives leave certain matters unprotected from the harmful effects of algorithmic bias. Therefore, while the EU anti-discrimination directives provide protection against algorithmic bias to a certain extent, it is limited in their scope. It is necessary to have further regulations that cover a wider range of areas to address algorithmic bias comprehensively.

5. Conclusion

The conclusion of this research paper aims to address the central question of whether European Union anti-discrimination law can provide a legal foundation for addressing the issue of algo-

rithmic bias. The research was divided into several sections, with each section building upon the previous one to provide a comprehensive analysis of the issue.

In the second section, the paper provided definitions for key terms, such as algorithms, machine learning, and automated decision-making, which were necessary for understanding the concept of algorithmic bias. This was followed by an examination of the EU anti-discrimination directives in section 3, which are the legal framework for combating discrimination in the EU. The directives prohibit direct and indirect discrimination based on certain protected grounds, including race, sex, religion, disability, age, and sexual orientation in employment, the welfare system, and access to goods and services. Section 4 explained how algorithmic bias can be classified as indirect discrimination under the EU anti-discrimination directives, and how it can be tested for proportionality. The argument was made that biased algorithms fail to satisfy the principle of necessity according to the principle of proportionality of Robert Alexy, and therefore, the discrimination caused by algorithmic bias is illegal under the EU anti-discrimination directives.

In conclusion, the paper acknowledges the discriminatory potential of algorithms but emphasizes that completely turning our back on their use is not the answer. Instead, it is essential to find ways to reduce their adverse effects. While the EU anti-discrimination directives offer a notable beginning to challenge algorithmic bias, they also have their limitations. As such, there is a need for forthcoming regulations that approach the subject in-depth and protect the fundamental rights of individuals while taking into account the needs of businesses. Further studies must examine how to regulate automated decision-making in a way that addresses the issue of algorithmic bias and promotes fairness and equality.

Acknowledgments

I would like to express my gratitude to the members of the Tarello Institute for welcoming me into their community. I appreciate the opportunity to be a part of such a prestigious institute and I look forward to growing and learning alongside its members. Additionally, I would like to acknowledge the use of OpenAI's ChatGPT language model for proofreading and grammar checking in this document.

References

- [1] A. Yapo, J. Weiss, Ethical implications of bias in machine learning (2018).
- [2] M. Veale, M. V. Kleek, R. Binns, Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making (pp. 1–14), 2018.
- [3] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (2016).
- [4] L. Kirchner, New york city moves to create accountability for algorithms, Ars Technica (2017).
- [5] The State of Global AI Regulations in 2023, Holistic AI, 2023.
- [6] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2017) 76–99.

- [7] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harv. JL & Tech.* 31 (2017) 841.
- [8] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a “right to explanation”, *AI magazine* 38 (2017) 50–57.
- [9] M. Kaminski, The right to explanation, explained (june 15, 2018). university of colorado law legal studies research paper no. 18-24, *Berkeley Technology Law Journal* 34 (2019).
- [10] T. H. Cormen, *Algorithms unlocked*, Mit Press, 2013.
- [11] R. Kitchin, Thinking critically about and researching algorithms, *Information, communication society* 20 (2017) 14–29.
- [12] E. Alpaydin, *Introduction to machine learning*, MIT press, 2020.
- [13] I. El Naqa, M. J. Murphy, *What is machine learning?*, Springer, 2015.
- [14] M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management, *Big Data Society* 5 (2018).
- [15] T. Araujo, N. Helberger, S. Kruikemeier, C. H. De Vreese, In ai we trust? perceptions about automated decision-making by artificial intelligence, *AI & society* 35 (2020) 611–623.
- [16] N. Diakopoulos, *Algorithmic accountability reporting: On the investigation of black boxes* (2014).
- [17] D. Pedreshi, S. Ruggieri, F. Turini, *Discrimination-aware data mining*, 2008, pp. 560–568.
- [18] M. Veale, R. Binns, *Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data*, *Big Data Society* 4 (2017).
- [19] B. Friedman, H. Nissenbaum, *Bias in computer systems*, *ACM Transactions on information systems (TOIS)* 14 (1996) 330–347.
- [20] E. Ellis, P. Watson, *EU anti-discrimination law*, OUP Oxford, 2012.
- [21] R. S. Baker, A. Hawn, *Algorithmic bias in education*, *International Journal of Artificial Intelligence in Education* (2021) 1–41.
- [22] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, *A survey on bias and fairness in machine learning*, *ACM Computing Surveys (CSUR)* 54 (2021) 1–35.
- [23] S. Barocas, M. Hardt, A. Narayanan, *Fairness and machine learning*. fairmlbook.org, 2019.
- [24] K. Lum, W. Isaac, *To predict and serve?*, *Significance* 13 (2016) 14–19.
- [25] J. Jung, R. Shroff, A. Feller, S. Goel, *Algorithmic decision making in the presence of unmeasured confounding* (2018).
- [26] J. Vanian, *Unmasking ai’s bias problem*, *Fortune* 25 (2018) 54–62.
- [27] A. Levendowski, *How copyright law can fix artificial intelligence’s implicit bias problem*, *Washington Law Review* 93 (2018) 579.
- [28] A. Howard, C. Zhang, E. Horvitz, *Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems*, *IEEE*, 2017, pp. 1–7.
- [29] K. Hee, *Is data quality enough for a clinical decision?: apply machine learning and avoid bias*, *IEEE*, 2017, pp. 2612–2619.
- [30] I. Zliobaite, *A survey on measuring indirect discrimination in machine learning* (2015).
- [31] T. Khaitan, *A theory of discrimination law*, OUP Oxford, 2015.
- [32] M. Minow, *Foreword: justice engendered*, *Harvard Law Review* 101 (1987) 10.
- [33] S. F. FBA, *Discrimination law*, Oxford University Press, 2011.
- [34] H. Collins, T. Khaitan, *Foundations of indirect discrimination law*, Bloomsbury Publishing, 2018.

- [35] L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, *Duke L. Technology Review* 16 (2017) 18.
- [36] E. T. Zouave, T. Marquenie, An inconvenient truth: algorithmic transparency accountability in criminal intelligence profiling, *IEEE*, 2017, pp. 17–23.
- [37] M. Borowski, Discourse, principles, and the problem of law and morality: Robert alexy's three main works, *Jurisprudence* 2 (2011) 575–595.
- [38] R. Alexy, Constitutional rights and proportionality, *Revus. Journal for Constitutional Theory and Philosophy of Law* (2014) 51–65.
- [39] R. Alexy, Balancing, constitutional review, and representation, *International journal of constitutional law* 3 (2005) 572–581.
- [40] F. Bonchi, S. Hajian, B. Mishra, D. Ramazzotti, Exposing the probabilistic causal structure of discrimination, *International Journal of Data Science and Analytics* 3 (2017) 1–21.
- [41] F. Kamiran, T. Calders, Classification with no discrimination by preferential sampling, Citeseer, 2010.
- [42] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, *Advances in neural information processing systems* 30 (2017).
- [43] J. Komiyama, H. Shimao, Two-stage algorithm for fairness-aware machine learning (2017).
- [44] C. A. Hidalgo, D. Orghian, J. A. Canals, F. De Almeida, N. Martin, How humans judge machines, MIT Press, 2021.
- [45] D. Morondo Taramundi, Discrimination by machine-based decisions: Inputs and limits of anti-discrimination law, in: *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, Springer, 2022, pp. 73–85.