Estudios de Deusto

Revista de Derecho Público

TECHNICAL FAIRNESS INTERVENTIONS IN AI SYSTEMS: A CRITICAL PERSPECTIVE

Intervenciones técnicas de equidad en sistemas de IA: Una perspectiva crítica

Anna Capellà i Ricart

https://doi.org/10.18543/ed.3748

Fecha de recepción: 15.07.2026 • Fecha de aceptación: 02.10.2026
Fecha de publicación anticipada en línea: noviembre 2026

Derechos de autores/as y Acceso Abierto

Se recomienda a las personas autoras que lean atentamente sus derechos. Creemos que este enfoque garantiza un acuerdo justo para ambas partes. Estas instrucciones se leen conjuntamente con el Código Ético y la Política de Acceso Abierto, Licencia y Derechos de Autor/a de la revista "Estudios de Deusto. Revista de Derecho Público".

1. Autoría y ética

Al entregar sus manuscritos a **Estudios de Deusto. Revista de Derecho Público,** los/las autores/as aceptan y se comprometen a cumplir las condiciones de publicación sin necesidad de firmar un documento de cesión adicional con la **Editorial (Universidad de Deusto).** Con ello, garantizan que su trabajo es inédito en cualquier forma, original y que no vulnera el Código Ético de la Revista ni

Copyright and Open Access

The authors are advised to read their rights carefully. We believe this approach ensures a fair agreement for both parties. These instructions should be read in conjunction with the Ethical Guidelines and the Open Access Policy, Licensing Terms and Copyright of the journal "Estudios de Deusto. Revista de Derecho Público".

1. Authorship and Ethics

By submitting their manuscripts to **Estudios de Deusto. Revista de Derecho Público**, the authors accept and undertake to comply with the conditions of publication without the need to sign an additional transfer agreement with the **Publisher (University of Deusto)**. In doing so, they guarantee that their work is unpublished in any form, original, and does not breach the Journal's Ethical Guidelines

© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
http://www.revista-estudios.deusto.es/

derechos de terceros, y que no se han otorgado ni se otorgarán licencias que resulten incompatibles con los derechos concedidos a la Editorial.

Los/las autores/as asumen la responsabilidad total y exclusiva sobre el contenido de su estudio y declaran formalmente no tener conflictos de interés que afecten la integridad de la investigación.

2. Derechos de la Editorial

Al enviar el manuscrito, los/las autores/as aceptan su publicación bajo la licencia **Creative Commons CC BY-NC-ND 4.0**. En consecuencia, conceden a la Editorial el **derecho exclusivo, gratuito y universal para la primera publicación**, edición, maquetación y explotación de la obra. Esta concesión autoriza a la Editorial a distribuir, sublicenciar e indexar el trabajo en cualquier formato, medio, base de datos o repositorio institucional, con fines de promoción y difusión científica.

3. Derechos de autores/as

Los/las autores/as **conservan la propiedad intelectual de su obra y retienen el derecho a distribuir y utilizar su trabajo** para fines docentes, investigación futura o archivo personal, siempre que se cite la publicación original en la Revista. Se les permite además la publicación posterior en otros medios, siempre que se incluya una nota a pie de página con la referencia completa de la Revista (incluyendo el DOI, si está disponible) y no se sugiera el respaldo explícito de la Revista y/o Editorial.

4. Acceso abierto

Estudios de Deusto. Revista de Derecho Público es una revista de acceso abierto; lo que significa que es de libre y total acceso en su integridad inmediatamente después de la publicación de su contenido. No obstante, de acuerdo con la licencia arriba mencionada, se debe citar siempre a los/las autores/as de los artículos. Tanto su uso comercial como cualquier modificación que se pretenda distribuir requerirán el permiso expreso previo por escrito de quien sea titular de los derechos.

or the rights of third parties, and that no licences have been or will be granted that are incompatible with the rights granted to the Publisher.

The authors assume full and exclusive responsibility for the content of the study and formally declare that they have no conflicts of interest that affect the integrity of the research.

2. Publisher's Rights

By submitting the manuscript, the authors agree to its publication under the **Creative Commons CC BY-NC-ND 4.0** licence. Consequently, they grant the Publisher the **exclusive, royalty-free and worldwide right of first publication**, editing, layout and exploitation of the article. This grant authorises the Publisher to distribute, sub-license and index the work in any format, medium, database or institutional repository, for the purposes of promotion and scientific dissemination.

3. Copyright

The authors **retain the intellectual property rights to their article and retain the right to distribute and use their work** for teaching purposes, future research or personal archiving, provided that the original publication in the Journal is cited. They are also permitted to republish in other media, provided that a (foot)note is included with the full reference to the Journal (including the DOI, where available) and no explicit endorsement of the Journal and/or Publisher is implied.

4. Open Access

Estudios de Deusto. Revista de Derecho Público is an open-access journal; this means that it is freely and fully accessible in its entirety immediately upon publication of its content. However, in accordance with the licence mentioned above, the authors of the articles must always be properly cited; and both commercial use and any modification intended for distribution will require the express prior written permission of the rights holder.

TECHNICAL FAIRNESS INTERVENTIONS IN AI SYSTEMS: A CRITICAL PERSPECTIVE

Intervenciones técnicas de equidad en sistemas de IA: Una perspectiva crítica

Anna Capellà i Ricart[1]
Profesora Lectora de Derecho Constitucional
Universidad Autónoma de Barcelona, Barcelona. España
https://orcid.org/0000-0002-9898-8242

https://doi.org/10.18543/ed.3748

Fecha de recepción: 15.07.2026
Fecha de aceptación: 02.10.2026
Fecha de publicación anticipada en línea: noviembre 2026

Abstract

AI systems can lead to discriminatory outcomes. While EU non-discrimination law prohibits discrimination on specific protected attributes, compliance of this legislation by AI decision-making systems implemented in social contexts appears uncertain for several reasons (e.g. AI systems often discriminate subtly and opaquely, making proof of discrimination difficult to obtain). This has led to the proliferation of technical fairness interventions designed to mitigate their discriminatory results. However, these technical fairness interventions are not socially discussed, and their objectives remain vague. In this article, we question whether such interventions are lawful under EU legislation and seek to clarify their precise objectives.

Keywords

Fairness, AI systems, affirmative action, EU anti-discrimination law, discrimination.

[1] Email: anna.capella@uab.cat

Resumen

Los sistemas algorítmicos pueden generar resultados discriminatorios. Mientras que el Derecho de la Unión Europea prohíbe la discriminación con base en atributos personales específicos, el cumplimiento de esta normativa por parte de los sistemas de decisión mediante inteligencia artificial en contextos sociales es incierto por diferentes motivos (por ejemplo, los sistemas de inteligencia artificial habitualmente discriminan de forma sutil y opaca, haciendo que la prueba de discriminación sea difícil de obtener). La incertidumbre en la normativa ha provocado la proliferación de intervenciones técnicas de equidad para mitigar los resultados discriminatorios de estos sistemas. Sin embargo, estas intervenciones no han sido objeto de discusión social y sus objetivos son vagos. En este articulo cuestionamos si estas intervenciones técnicas de equidad son legales en relación con el marco legal de la Unión europea e intentamos clarificar cuales pueden ser sus objetivos.

Palabras clave

Equidad, sistemas de IA, acción positiva, normativa antidiscriminación europea, discriminación.

Summary: I. Introduction. II. Discriminatory outcomes in AI systems. III. Technical fairness interventions in AI systems. IV. EU legal framework: anti-discrimination law, GDPR and AI act. 1. Article 10(5) of the AI Act. 2. An affirmative action measure? V. What is the aim of technical fairness interventions? VI. Conclusions. VII. References.

I. INTRODUCTION

To tackle algorithmic discrimination, the EU has established a legal framework of non-discrimination laws, the General Data Protection Regulation (GDPR), and the Artificial Intelligence Act (AI Act). While these regulations can address some problems caused by algorithmic discrimination, they also present certain drawbacks and inconsistencies.

Each of these regulations adopts a different perspective regarding the meaning of 'fairness', a fact that can impact the development of technical fairness interventions. Recital 27 of the AI Act states that, for the purpose of the regulation, 'diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law'. Holtz and Ledendal (2026, 15) highlight that, in the AI Act, the principle of fairness must be understood in conjunction with the principle of accuracy. In the GDPR, however, fairness appears as a principle of data processing alongside the principles of lawfulness and transparency, requiring that personal data should not be processed in a way that is unjustifiably detrimental, unlawfully discriminatory, unexpected, or misleading to the data.[2] From a social sciences perspective, the multiple definitions of fairness (Calvi and Kotzinos 2023, 1229) extend beyond the definitions or guidelines established in the AI Act and the GDPR (which address fairness not in isolation but in conjunction with other principles).

Two of the principal problems with the current legal framework are: i) the regulations do not provide concrete standards that describe the characteristics of a fair AI system; and ii) although affirmative action measures are recognised in EU anti-discrimination law, they are restricted when they lead to what is commonly termed 'reverse discrimination'.[3]

[2] EDPB Guidelines 4/2019 on Article 25 — Data Protection by Design and by Default, Version 2.0, adopted 20 October 2020, paragraph 69. https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf, visited 11 May 2026.

[3] See, for instance: ECJ 17 October 1995, Case C-450/93, *Eckhard Kalanke* v *Freie Hansestadt Bremen.* From an anti-subordination perspective, as we will now address

In this context, technical fairness interventions applied to algorithms (e.g. overall accuracy equality, predictive parity, or fairness through unawareness), fairness-related datasets, and fairness-enhancing mechanisms, primarily developed by engineers, have emerged (Pessach and Shmueli 2022, 23).

However, there is a lack of guidance and debate concerning the priorities and objectives of these interventions. As for priorities, several possibilities exist: should the focus be on individual or group fairness? Should biased data be modified, fairness constraints added directly to a model's objective function, or a biased model's outcome adjusted?

Various alternatives also present themselves regarding their objectives. One aim could be to ensure that AI systems do not amplify or magnify existing discrimination (e.g. creation of a digital representation of an already discriminatory society). However, the extent to which this is sufficient is debatable. As Chouldechova and Roth (2020, 87) affirm, 'the observed world is not one we should seek to replicate accurately in the first place'.

The goal could also be to make algorithmic systems less discriminatory compared to current social discrimination. Yet, who determines the extent to which AI systems must be 'less discriminatory'? And on what standards should these principles be built? While applying a technical intervention may improve the treatment given to a particular social group by an algorithm, it could also potentially aggravate the treatment received by another social collective that, without the intervention, would have received a more beneficial outcome. Does this constitute reverse discrimination? Moreover, affirmative action measures may be permitted in some countries but not in others, which could be problematic given that AI systems lack geographical boundaries.

Lastly, the possibility that these interventions aim to enhance the perception that companies are striving to make algorithms fairer (if possible) solely to justify their use cannot be disregarded. In this regard, the intervention could be viewed as 'fairness washing' — understood as feigning consideration of discriminatory impacts to improve how AI systems are perceived.

It is widely believed that technical fairness interventions can influence discriminatory outcomes. However, it is important to review and adjust this affirmation to more realistic expectations and to pose crucial questions.

The purpose of this paper is to approach the implementation of technical interventions in algorithmic systems from a legal perspective, moving beyond purely technical studies. The paper aims to analyse whether technical fairness interventions applied to AI systems (other than those implemented according to Article 10(5) of the AI Act) are lawful under EU law.

---

more thoroughly, there is no such thing as reverse discrimination because discrimination can only occur when affects people from social groups that find themselves stigmatised, subordinated or excluded.

The research methodology is based on reviewing and analysing relevant law, jurisprudence, and legal doctrine. Building on this methodology, we will first briefly explain why AI systems can have discriminatory outcomes. Next, we will define and study technical fairness interventions in algorithmic systems. Hereafter, we will present the main aspects of EU regulation on algorithmic discrimination. Then, we will reflect on the aim of technical fairness interventions. Lastly, we will present some conclusions.

II. DISCRIMINATORY OUTCOMES IN AI SYSTEMS

Unfairness in AI systems can arise from issues within datasets, such as unrepresentative or decontextualised data, or the introduction of human bias during data labelling. Additionally, algorithmic functions can contribute to unfairness; for example, the capacity of algorithmic systems to infer protected attributes from data not explicitly representing them (e.g. inferring ethnic origin from a post code, gender from part-time work, or ethnic origin from a name). Furthermore, human perceptions introduced into algorithmic systems, such as bias in data labelling or subjectivity in variable classification, can result in unfairness. Finally, humans may instrumentalise these possibilities to mask discriminatory motives or deploy algorithmic decision-making systems despite their known discriminatory outcomes,[4] as seen in the case of the Austrian Public Employment Service (Allhutter et al. 2020) or the Spanish BOSCO algorithmic system (Belloso Martín 2022, 49).

From a broader perspective, we live in an inherently unequal society where discriminatory actions influence social interactions and decisions.[5] In this context, unfairness in AI systems can be viewed either as a specific error or deviation by the machine (as outlined in the first paragraph) or as a replication of our discriminatory society. The latter carries the potential to magnify and perpetuate societal discrimination and, crucially, to stifle efforts towards greater equality. These perspectives are linked to differing approaches in understanding the aim of anti-discrimination law, a discussion that impacts how the EU addresses discrimination in AI.

From an anti-classification perspective of anti-discrimination law, discrimination occurs when arbitrary differences in treatment are established based on morally irrelevant criteria. Not all arbitrary differences in treatment

[4] The classification presented in this paragraph derives from the following paper: Barocas and Selbst 2016, 671.

[5] From our perspective, this is one of the main lessons we can learn from discriminatory AI systems: these systems learn from data and its correlations and therefore, mainly, their discriminatory results are the translation of discriminatory patterns to machines. What causes discrimination is more a system than an error.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

are discriminatory; only those that involve protected attributes of individuals. These protected attributes (e.g., sex, ethnic origin, age, sexual orientation) should not be relevant in decision-making. Therefore, this perspective equates non-differentiation with equality. Consequently, an algorithm is deemed discriminatory when it considers a protected attribute in making a decision, as this attribute should have been irrelevant.

Taking an anti-subordination view, discrimination is defined as action against individuals belonging to social collectives that are historically and currently stigmatised, subordinated or excluded. It is not the treatment of an individual, but the treatment of a group of people (subject to prejudices, stereotypes and pejorative treatment) who share one or several protected attributes that causes algorithmic systems to generate discriminatory results. This applies whether the result is individualised (harming a specific person) or collective (harming the entire group). Examples include AI systems reproducing gender stereotypes or being used to undress women and distribute the content.[6] This perspective highlights that ignorance of group differences can perpetuate inequality instead of eradicating it. Ho and Xiang (2020, 134) argue that if bias mitigation is not linked to historical discrimination (empirically documented and attributable to the deploying entity), algorithmic fairness may reach a dead end.

III. TECHNICAL FAIRNESS INTERVENTIONS IN AI SYSTEMS

Biases may ultimately stem from the context in which the AI system is deployed, rather than solely from the data used to train it. Nevertheless, removing biases from datasets and the algorithm itself can be a valuable starting point. Technical fairness interventions have increased rapidly (Pessach and Shmueli 2022, 23), some examples being the use of fairness-aware datasets (Mehrabi et al., 2021, 1), equalised odds (Hardt et al., 2016, 3315), or predictive parity.[7] These interventions can target audiences, individuals or social groups and can be implemented at different stages of the algorithmic system's design and deployment.

Regarding target audiences, algorithmic individual fairness aims to ensure that an algorithm's decision is influenced only by an individual's relevant characteristics, not by group membership. In other words, it seeks to ensure that the algorithm provides similar predictions for similar individuals.

[6] We consider this harm collective because the female's body is devalued and sexually objectified. However, this harm could also be considered individual (regarding the woman who has suffered the harm).

[7] For a catalogue of different technical fairness interventions, see: Burda and Van Otterloo, 2025; Castelnovo et al., 2022; Caton and Haas, 2024.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

Dwork (2010, 214) posits that similarity is defined with respect to a task-specific metric, which must be determined on a case-by-case basis, presupposing the existence of an agreed-upon similarity metric. Fairness, understood as 'the quality of treating people equally or in a way that is right or reasonable'[8] is considered here between individuals: as the quality of an algorithm to treat one individual equally to another. This perspective aligns with the anti-classification approach to anti-discrimination. Anderson and Visweswaran (2025, 1) highlight several limitations of this approach, such as the difficulty in defining 'similar individuals' and the fact that relying on learning similarity metrics from data can embed existing biases within that data.

Conversely, algorithmic group fairness aims to achieve parity of statistical metrics across groups that differ in one or several protected attributes (e.g. male and female groups). This involves searching for properties that perform equally well across protected groups (i.e., that all groups have similar false positives rates). As Hellman (2020, 811) points out, 'while a lack of parity in this dimension is not constitutive of unfairness, this measure provides important reasons to suspect that unfairness exists'. Fairness, in this context, is the quality of an algorithm to treat social groups of people equally compared to other social groups; it is appreciated with respect to the whole group. However, these technical interventions, which focus on social collectives, face the problem of the existence of multiple groups of people with overlapping memberships or complex relationships (Anderson and Visweswaran 2025, 1). If all protected attributes under EU law are merged into one group of people belonging to a subordinated or excluded social group regarding sex, ethnic origin, disability, age, sexual orientation, and religion or belief, who will be the comparator group to assess the effectiveness of this intervention? Chouldechova and Roth (2020, 85) raise a similar point when they affirm that this kind of technical interventions 'give[s] guarantees to "average" members of the protected groups', not, for example, those who simultaneously belong to different protected groups.

Moreover, interventions can be classified according to the moment of their application. 'Pre-processing algorithms' aim to modify biased data (e.g. adjust, transform, reweight, or augment data to ensure balanced representation) (Zhang et al., 2023). 'In-processing algorithms' seek to add a fairness constraint directly to a model's objective function (Kearns et al., 2019), and 'post-processing algorithms' aim to modify biased model outcomes (e.g. adjust or transform the decisions or outputs of an algorithm after its training) (Petersen et al., 2021; Small et al., 2024; Nabi and Shpitser, 2018).

[8] Fairness. Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/fairness, visited 11 May 2026.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

Nonetheless, if the algorithm uses biased training datasets, the extent these interventions are useful is debatable.

The researchers who develop these fairness interventions acknowledge that their implementation implies certain trade-offs. One example is the COMPAS algorithm, used to measure the recidivism risk score of offenders (as if this could be done ethically): although it had the same equal accuracy of positive predictions for different ethnic groups (equalised accuracy), people from these different ethnic groups did not have equal odds of receiving a positive prediction (equalised odds).

Furthermore, fairness may come at a cost to other algorithmic properties, such as accuracy. Adding features is commonly believed to enhance the accuracy of predictions for all groups, and therefore the erasure or unavailability of data concerning protected attributes is sometimes thought to be detrimental to all groups (Kleinberg et al., 2017; Pessach and Shmueli 2022). Conversely, Höltgen and Oliver (2025, 241) and Yao et al., (2024, 1) claim that fairness through unawareness can reduce algorithmic discrimination without necessarily reducing accuracy.

In any case, it has been pointed out that all forms of discrimination cannot be avoided when making decisions based on statistics (Kleinberg et al., 2017). Attention has also been drawn to the fact that combining multiple fair components often fails to achieve overall fairness (Bower et al., 2019, 1; Dwork and C. Ilvento 2019, 1).

Moreover, Pessach and Shmueli (2022, 23) question how the effectiveness of each technical intervention can be evaluated and which mechanisms should be used for comparison. The issue raised by these authors leads us to further questions: i) To what extent are these interventions effective when the algorithm evolves (or keeps learning)? ii) Do these interventions have limited (and also unexpected) effects? iii) If the interventions proposed by specialised doctrine only impact a specific decision point, are the structural causes and effects of discrimination considered? iv) Are these interventions conceptualised considering interdisciplinary perspectives? v) Are they legal under EU law?

IV. EU LEGAL FRAMEWORK: ANTI-DISCRIMINATION LAW, GDPR AND AI ACT

EU law provides various instruments to address discrimination caused by algorithmic systems. Firstly, Article 21 of the Charter of Fundamental Rights of the European Union (CFREU) prohibits discrimination based on any ground. Article 19 of the Treaty on the Functioning of the European Union (TFEU) establishes that the Council may take appropriate action to combat discrimination based on sex, racial or ethnic origin, religion or belief,

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

disability, age or sexual orientation. Based on this article, a body of EU secondary law has emerged to combat discrimination: Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services; and Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast), among others. These directives aim to ensure the effective implementation of the principle of equality and to protect all natural persons against discrimination on the grounds previously referred to, adapting means of legal protection such as rules on the burden of proof.

These directives specifically permit member states, with a view to ensuring full equality in practice, to maintain or adopt specific measures to prevent or compensate for disadvantages linked to any protected attribute (Article 5 Directive 2000/43/EC; Article 7.1 Directive 2000/78/EC; Article 6 Directive 2004/113/EC; Article 3 Directive 2006/54/EC). In EU law, affirmative action is only expressly considered for these protected attributes.

The legal framework established in these directives permits the justification of discrimination. For example, direct discrimination can be justified by reason of the nature of the particular occupational activities concerned (Article 4, Directive 2000/43/EC), and indirect discrimination by objective justification. Discrimination is understood as an isolated phenomenon (i.e. someone has wrongly made a decision, creating a difference of treatment between people) and not as an epiphenomenon of a situation of stigmatisation, subordination or exclusion. Furthermore, the directives have different scopes of application according to the protected attribute they intend to protect (e.g. the labour market; welfare and social protection; access to goods and services). Lastly, even if there are rules to reverse the burden of proof when a *prima facie* case of discrimination has been established, the presentation of evidence remains difficult when algorithmic systems are involved.[9] This is due to several factors: i) Algorithmic systems can be opaque, making it difficult to ascertain if discrimination has occurred or to find comparable

[9] These issues regarding the burden of proof could have been addressed by the European Commission proposal on the AI Liability Directive that aimed to introduce harmonised rules on the non-contractual liability for damage caused by artificial intelligence; however, the European Commission announced its withdrawal in its 2025 work programme.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

cases involving other people; ii) Proving responsibility for discrimination is notoriously challenging, given the multitude of companies and individuals involved in the process of creating an algorithmic decision-making system (i.e. companies, organisations or individuals that design, develop or implement, establish objectives for, make decisions related to, ensure the operation of, or use the system); iii) The legal framework and the European Court of Justice (ECJ) barely recognise the existence of an intersectional perspective of discrimination, thereby reducing the analysis of cases to one protected attribute at a time;[10] iv) The abstract nature of profiling can make it impossible to identify the exact attributes that lead to discrimination.

As decision-making through automated means can involve the processing of personal data, the GDPR is also applicable in cases of algorithmic discrimination. Several articles in the regulation can be highlighted to address this issue and the lack of information surrounding it: Article 9 (prohibition to process special categories of data); Article 22 (right not to be subject to solely automated decision-making); Article 5(d) (data accuracy); Articles 15-22 (data subject's rights); and Articles 13(2)(f), 14(2)(g) and 15(1)(h) (accessing information on the algorithmic decision-making process) in relation to Recital 71 (automated processing should include specific information to the data subject, and technical and organisational measures must be implemented to correct inaccuracies).

Nonetheless, this regulation has certain drawbacks when approaching discriminatory outcomes of algorithmic decision-making. For instance, big data analysis can infer sensitive data from sources other than those classified as "special category" (Van Bekkum and Zuiderveen Borgesius 2023, 1; Žliobaitė and Custers 2016, 183; Kroll et al., 2017, 685; McLoughney et al., 2023, 1). Moreover, while data subjects' rights can help improve data accuracy and prevent discriminatory results, the effectiveness of these rights in the context of big data analytics is contentious. Aslam (2025, 362) questions the right to erasure in machine learning scenarios, noting that data cannot always be completely erased and that erasure can have varying consequences for the overall system, raising ethical implications. Furthermore, the existence and effectiveness of the right to explanation under the GDPR have been widely debated (Metikoš and Ausloos 2025, 205; Bayamlıoglu 2022, 1058; Kaminski 2019, 189; Wachter, Mittelstadt, and Floridi 2017, 76). The ECJ recently clarified that the 'GDPR requires the controller to provide meaningful information about the logic involved in [the] decision, but "not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm"', specifying that the 'controller should find simple ways to tell

[10] This is counter-productive in algorithmic discrimination because, thanks to the ability to correlate data, discrimination will rarely occur based on a single protected attribute.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

the data subject about the rationale behind, or the criteria relied on in reaching the automated decision'.[11]

Lastly, Article 1 of the AI Act states that one of its primary objectives is to ensure a high level of protection for fundamental rights, such as the right to equality and non-discrimination, which is enshrined in article 21 of the CFREU. The Act establishes protection based on the risk an AI system may pose. Several articles within the Act are pertinent to tackling algorithmic discrimination, particularly in high-risk AI systems. Article 9 mandates the establishment, implementation, documentation and maintenance of a risk management system. Article 10 sets out obligations regarding data quality and accuracy, with Article 10(5) specifically enabling providers of high-risk AI systems to process special categories of data to detect and correct biases. Article 11 establishes the obligation to draw up technical documentation, which details on potential discriminatory impacts (Annex IV(2)(g)) that can help individuals prove discriminatory outcomes suffered due to specific AI systems. Article 12 states the obligation of record-keeping of events. Article 13 establishes obligations regarding transparency, and Article 14 mandates the existence of human oversight.

Furthermore, Article 27 obliges a fundamental rights impact assessment for high-risk AI systems to evaluate the impact their use may produce. Article 57 establishes AI regulatory sandboxes aimed at, among other things, improving legal certainty to achieve regulatory compliance, thereby making it possible to anticipate and detect discriminatory outcomes. Article 86 states the right to obtain clear and meaningful explanations from the deployer regarding the AI system's role in the decision-making process and the main elements of the decision. Additionally, certain information (Annex VIII) pertaining to high-risk AI systems must be made available in a public database (Article 71). Chapter IX of the AI Act (Articles 72 to 94) establishes a system of post-market monitoring.

However, this Act presents certain shortcomings. While some AI systems are prohibited, discrimination is generally perceived as an assumable risk and the regulation does not establish concrete results (e.g. accuracy rates in relation to different population groups, considering combined categories, or the obligation to prove that the system does not magnify existing structural discrimination). Moreover, providers are primarily responsible for conducting self-assessments of high-risk AI systems' requirements, with some exceptions where this task is attributed to notified bodies. Although a right to explanation has been clearly affirmed, its effectiveness in discrimination cases remains ambiguous. As per the judgment of the Spanish Supreme Court 1119/2025, of

[11] ECJ 27 February 2025, Case C-203/22, *CK* v *Magistrat der Stadt Wien*, paragraph 60.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

11 September,[12] explanation is an effective guarantee against arbitrariness and discriminatory biases. Nevertheless, and this argument also applies to the GDPR's conception of the right to an explanation, individuals must know what to ask for and how to interpret the information provided by deployers. They must also have the time and interest to engage in this kind of activity and learn the channels and processes to pursue to obtain the information.

In the current, fragmented legal framework concerning algorithmic discrimination, technical fairness interventions have emerged. It is imperative to examine whether their implementation is permissible under EU law, specifically Article 10(5) of the AI Act or as affirmative action measures.

1. *Article 10(5) of the AI Act*

Legally distinct situations arise when a fairness intervention corrects bias (modifying estimates that are notably inconsistent with true figures), merely prevents discrimination (advocating for a more equitable society through awareness, institutional policies and inclusive behaviours), or provides preferential treatment to a protected or disadvantaged group. It must be acknowledged that, due to the opacity of certain AI systems, identifying which of these scenarios is occurring can be challenging.

The only specific mention in EU law of any measure to prevent or compensate for disadvantages linked to protected attributes in AI systems is found in Article 10(5) of the AI Act. This article represents the first attempt by any legislator worldwide to balance the fundamental rights to data protection and non-discrimination within the context of AI debiasing (Van Bekkum 2025, 11).

This article permits providers of high-risk AI systems to process special categories of personal data exceptionally, to the extent strictly necessary for bias detection and correction, in accordance with Article 10(2)(f)[13] and (g).[14] Such processing is subject to appropriate safeguards for the fundamental rights and freedoms of natural persons and must comply with specific conditions.[15]

---

[12] Spanish Supreme Court, 11 September 2025, 1119/2025.

[13] Training, validation and testing datasets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: (f) 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.

[14] Training, validation and testing datasets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f).

[15] The conditions include: a) no availability of other data to fulfil the purpose; b) no re-use and implementation of security and privacy-preserving measures; c) controls and

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

Van Bekkum explains that the legislator has limited the scope of this article to bias detection and correction, excluding bias monitoring (meaning special categories of data cannot be used once the AI system is operational). Furthermore, the exception applies solely to providers of high-risk AI systems, not third parties, thereby restricting the storage and access to such data (Van Bekkum 2025, 10).

Article 10(5) of the AI Act also constitutes an exception to Article 9 of the GDPR, which prohibits the processing of certain special categories of data. This exception addresses warnings from specialised doctrine regarding the ability of big data analysis to infer special categories of data from other information (Van Bekkum and Zuiderveen Borgesius 2023, 1; Žliobaitė and B. Custers 2016, 183; Kroll et al., 2017, 685; McLoughney et al., 2023, 1). If these special categories of data were unavailable to the algorithm, it could infer the same information from other data, making bias more difficult to identify, detect and correct.

Nonetheless, Ho and Xiang (2020, 134) caution that the use of protected attributes should not be taken lightly, as it is likely to amplify rather than mitigate bias. For instance, the authors cite the example of race in medicine, where the use of race variables in clinical diagnostic tools is often deemed acceptable due to the perception that race is genetically based. However, this approach can lead to the misattribution of health outcome differences to genetic rather than social factors.

Article 10(5) of the AI Act permits the implementation of specific fairness interventions under strict conditions. From an anti-discrimination law perspective, the imposition of strict conditions reflects the often blurry line between permissible affirmative action and discrimination within the current legal framework. For example, Hoch et al., (2024, 13) analysed an algorithm that incorporated a fairness constraint regarding sex. They concluded that on the one hand, the use of protected attributes could be considered permissible as an affirmative action under EU law for the sake of achieving a fairer outcome. On the other hand, however, they argued that *prima facie* arguments for indirect discrimination against women could arise, as well as potential claims of direct discrimination against men.

Beyond Article 10(5), the legislation establishes a broader governance structure, encompassing obligations related to data governance, risk management, transparency, human oversight, accuracy, fundamental rights impact assessment and post-market monitoring. This governance structure aims, among other objectives, to ensure a high level of protection for fundamental rights enshrined in the CFREU against the harmful effects of AI systems.

documentation of access; d) no transmission, transfer or access by third parties; e) deletion after the bias is corrected; and f) records that include the reasons why the processing was necessary and not achievable with other data.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

This implies that interventions in AI systems designed to enhance their respect for fundamental rights can and must be implemented. Notwithstanding this, we argue that it remains unclear to what extent interventions implemented in AI systems, beyond the detection and correction of bias, are permitted under EU law from an affirmative action and reverse discrimination perspective.

2. *An affirmative action measure?*

As we have observed, EU anti-discrimination law allows for member states to adopt specific measures to prevent or compensate for disadvantages linked to any protected attribute. Examples of affirmative action measures include the implementation of gender or disability quotas in recruitment processes, campaigns to promote Science, Technology, Engineering, and Mathematics (STEM) careers among girls (and, though not yet implemented, the promotion of 'caring' careers among boys, considering this measure could help prevent disadvantages linked to sex), or the prioritisation of women or individuals from specific ethnic origins in under-represented sectors.

It is clear that the implementation of affirmative action measures necessitates taking into account the protected attribute to deploy the measure, as opposed to an anti-classification perspective where the goal is for the protected attribute not to be considered in decision-making.

The deployment of affirmative action measures has caused some controversy when leading to reverse discrimination. From an anti-subordination perspective, the term 'reverse discrimination' is a misnomer, given that the legal meaning of 'discrimination' applies exclusively to individuals belonging to excluded, subordinated or stigmatised social groups, and the prejudice resulting from such unequal treatment is of a different nature — not discriminatory (Álvarez del Cuvillo 2022, 190-191). Reverse discrimination can be defined as treatments that cause harm to people who belong to dominant or majority social groups precisely because of their membership in these groups (Fredman 2011, 166). Individuals who experience 'reverse discrimination' do not feel a duty (neither moral nor legal) to bear the consequences of measures established to combat group discrimination.

In the context of EU law, the ECJ has addressed controversial affirmative action measures primarily concerning sex discrimination.[16] These measures aimed to establish selection criteria to balance competition and afford women a greater chance of obtaining a specific job than they would have had without

[16] In the United States of America, mainly regarding ethnic discrimination. *See*: *Ricci v. DeStefano*, 557 U.S. 557 (2009) and *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*, 600 U.S. (2023).

such intervention (see: Kalanke,[17] Marschall,[18] Badeck,[19] and Abrahamsson[20]). The ECJ ruled that affirmative action measures in a competitive employment context were invalid if they granted automatic preference to female candidates.[21] A job could be awarded to a female candidate if there was equivalence in the merits of the candidates —meaning that while the merits need not be identical, there could not be a considerable difference[22]— and provided an opening clause existed[23] (i.e. preference unless there were reasons for the male candidate to tip the balance in his favour). Similarly, other cases where affirmative action measures granted advantages to women were discussed, consistently treating them as an exception to formal equality of treatment (see: Lommers,[24] Briheche,[25] Griesmar,[26], Alvarez,[27] and Leone[28]).

ECJ case law shows that affirmative action must satisfy a proportionality test (Nowak 2022, 477; Waddington and Bell 2011, 1506), with greater emphasis on the legitimacy of the measure than on its effectiveness or necessity (Nowak 2022, 477). Although EU law does not explicitly refer to the temporary nature of such measures, it is understood that this is a factor to be considered when determining their proportionality (Waddington and Bell 2011, 1511). Furthermore, the measure must remain within the limits of what is appropriate and necessary to achieve the stated objective, requiring the

[17] ECJ 17 October 1995, Case C-450/93, *Eckhard Kalanke* v *Freie Hansestadt Bremen.*

[18] ECJ 11 November 1997, Case C-409/95, *Hellmut Marschall* v *Land Nordrhein-Westfalen*.

[19] ECJ 28 March 2000, Case C-158/97, *Badeck and Others*, *interveners: Hessische Ministerpräsident and Landesanwalt beim Staatsgerichtshof des Landes Hessen.*

[20] ECJ 6 July 2000, Case C-407/98, *Katarina Abrahamsson and Leif Anderson* v *Elisabet Fogelqvist*.

[21] ECJ 17 October 1995, Case C-450/93, *Eckhard Kalanke* v *Freie Hansestadt Bremen*, § 22.

[22] ECJ 6 July 2000, Case C-407/98, *Katarina Abrahamsson and Leif Anderson* v *Elisabet Fogelqvist*, § 62.

[23] ECJ 11 November 1997, Case C-409/95, *Hellmut Marschall* v *Land Nordrhein-Westfalen*, § 33.

[24] ECJ 19 March 2002, Case C-476/99, *H. Lommers v Minister van Landbouw, Natuurbeheer en Visserij*.

[25] ECJ 30 September 2004, Case C-319/03, *Serge Briheche v Ministre de l'Intérieur, Ministre de l'Éducation nationale et Ministre de la Justice*.

[26] ECJ 29 November 2001, Case C-366/99, *Joseph Griesmar v Ministre de l'Economie, des Finances et de l'Industrie and Ministre de la Fonction publique, de la Réforme de l'Etat et de la Décentralisation.*

[27] ECJ 30 September 2010, Case C-104/09, *Pedro Manuel Roca Álvarez v Sesa Start España ETT SA*.

[28] ECJ 17 July 2014, Case C-173/13, *Maurice Leone and Blandine Leone v Garde des Sceaux, ministre de la Justice and Caisse nationale de retraite des agents des collectivités locales.*

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

reconciliation, as far as possible, of the principle of equal treatment with the requirements of the objective pursued.[29] The Court here implies that the means employed must be appropriate for the objective sought, and that interests must be weighed such that the consequences are justified based on the importance of that objective.

The ECJ interprets lawful affirmative action measures very narrowly. Consequently, technical fairness interventions in AI systems may be contested if their implementation benefits certain individuals or groups at the expense of others (who would have received more favourable treatment from the algorithm without the intervention). In such cases, *prima facie* arguments for reverse discrimination could arise. Nevertheless, the opacity of some systems can prevent the identification of whether interventions have favoured particular groups.

Furthermore, for technical fairness interventions in AI systems to be considered an affirmative measure, they must be appropriate, necessary and proportionate, while also being reconciled, as far as possible, with the principle of equal treatment (as formally understood).

However, these requirements (especially the latter) are open to various interpretations, which are difficult to assess and discuss in contexts involving AI systems. For instance, in the ECJ judgment of 15 May 2025, *UV and XXX v Instituto Nacional de la Seguridad Social (INSS)*,[30] the court precluded member states from implementing a measure that, 'with a view to reducing the gender gap in connection with social security benefits which is attributable to the bringing-up of children, a pension supplement is awarded to women who are in receipt of a contributory retirement pension and who have had one or more children, whereas the award of that supplement to men placed in an identical situation is subject to additional requirements relating to whether their professional career was interrupted or affected by the birth or adoption of their children' (paragraph 77). The judgment specifies that men can be in an identical situation to women when raising children. The conflict in this case (which can be extrapolated to other affirmative action measures) is as follows: an affirmative action measure focuses on assisting women, as they collectively undertake more caring obligations, impacting their labour expectations and retirement pensions. However, disregarding the social impact of this task, individual men do not feel a duty (neither moral nor legal) to bear the consequences of measures established to combat group discrimination in their personal cases. The problem here lies in the fact that

[29] ECJ 19 March 2002, Case C-476/99, *H. Lommers v Minister van Landbouw, Natuurbeheer en Visserij*, § 39.

[30] ECJ 15 May 2025, Joined Cases C-623/23 and C-626/23, *UV and XXX* v *Instituto Nacional de la Seguridad Social (INSS)*.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

even though these men do not feel this duty, the law allows for the adoption of specific measures to prevent or compensate for disadvantages linked to any protected attribute (women who historically and currently perform caregiving tasks that impact their labour prospects and conditions). The ECJ case-law, however, limits these measures, arguing that they should be reconciled, as far as possible, with the principle of equal treatment (meaning that everyone —men and women— is treated the same).

Weerts et al. (2024, 2060) have proposed that algorithmic fairness interventions should often be interpreted as a means to prevent discrimination, rather than as measures of affirmative action. The authors argue that considering fairness interventions in algorithmic systems as affirmative actions would impose a significant legal burden on AI users, owing to the ECJ strict conditions for assessing the lawfulness of affirmative action measures. Moreover, the authors suggest a more proactive approach —a positive obligation to actively do no harm— as a more adequate framework for algorithmic decision-making.

However, it is complex to argue that technical fairness interventions are positive obligations to actively 'do no harm' rather than affirmative action measures, based on the premise that algorithmic decision-making is neither neutral nor objective. This is because the same argument could be extrapolated to human decision-making, which is neither neutral nor objective in the same contexts. What distinguishes algorithmic decision-making, however, is its capacity to impact a far greater number of people and to amplify existing discrimination.

Therefore, it is difficult to justify different applications of the law concerning equality, non-discrimination or fairness for human and algorithmic decision-making processes. Such differentiation would only be justified if the aim is to mitigate algorithmic decision-making's ability for 'amplification', as this exacerbates discrimination. As Chander (2017, 1045) points out, to prevent viral discrimination (understood here as synonymous with this 'amplification ability'), affirmative action principles must be applied.

Chouldechova and Roth (2020, 82) affirm that 'our understanding of the fundamental questions related to fairness and machine learning remain in its infancy'. The authors pose the question: what should fairness mean? From our understanding, fairness should not have a different meaning when related to machine learning. The problem is that performing interventions on machine learning requires a clear definition of fairness, and indeed, the meaning of 'fairness' remains a subject of ongoing debate. As a society, if we wish to implement technical fairness interventions in algorithmic systems while we remain perpetually discussing the meaning of 'fairness', what are our options? i) Not to use algorithmic systems; ii) Address only the additional problems that algorithms pose to social inequalities and discrimination.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

To summarise, it does not seem valid to justify a positive obligation to actively 'do no harm' when developing algorithmic decision-making systems if this positive obligation does not exist in human decision-making systems. The current EU framework neither includes nor forbids the possibility of implementing technical fairness interventions in algorithmic systems. Moreover, these interventions can impact the perception of systems as fairer, when, in actuality, it is unclear how the system is fairer, based on what standards, and compared to what systems. Should these interventions be implemented then, without clear guidance on what is expected from them?

V. WHAT IS THE AIM OF TECHNICAL FAIRNESS INTERVENTIONS?

The development of technical fairness interventions within complex and ambiguous legal frameworks can lead to the following questionable assertions, such as: 'it seems that biased algorithms are easier to fix than biased humans or procedures' (Pessach and Shmueli 2022, 24) or more easily 'eliminate existing discrimination' (Kleinberg et al., 2018, 164). Firstly, it is improbable that algorithms will self-correct biases or agree on how to address them in social interactions if humans fail to do so themselves.[31] Therefore, without a fundamental change in human behaviour, it is implausible that algorithms will change with respect to discriminatory outcomes (beyond changes resulting from achieving an equivalence between societal and algorithmic discrimination). Secondly, how can discrimination be eradicated from algorithms if it cannot be eradicated from social interactions?

This prompts an examination of the precise aim of technical fairness interventions. As previously stated, these interventions are not explicitly enabled by the EU legal framework, with the exceptions of processing special categories of data under Article 10(5) of the AI Act or affirmative action measures under specific EU directives.

From our perspective, there are three possible objectives. Firstly, the aim of these interventions might be for algorithmic systems to be no more discriminatory than society itself, thereby implicitly affirming that society is sufficiently equitable as it stands. In this scenario, technical fairness interventions would be implemented to achieve two goals: firstly, to be as accurate as human decision-making concerning individuals and social groups; and, secondly, to mitigate the 'amplification' capability of algorithmic systems. This would imply that algorithms are equally accurate as humans in reflecting structural, systemic and collective discriminatory perceptions.

[31] In the following lines, however, we will address the possibility that fair representation learning in algorithms impact the promotion of fairness in the analogue world.

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

If technical fairness interventions had this objective, as Weerts et al., point out, they could or could not be considered affirmative action measures. On the one hand, they would not fit the definition, as their objective is not to prevent or compensate for disadvantages linked to any protected attribute, but to ensure that algorithmic decision-making systems are no more discriminatory than human decision-making systems. On the other hand, they could fit into the legal category because, even though their primary aim is not to improve the quality of life for certain collective social groups, this could, in fact, be a result of their implementation. However, the outcome of their implementation is unclear, given the multitude of factors to consider and the fact that, as we have stated, applying a technical fairness intervention does not guarantee that all forms of discrimination can be avoided.

Secondly, the objective of these interventions could be to make algorithmic systems less discriminatory than society. For example, to achieve gender representation in the digital environment free of gender roles and stereotypes (though much work is needed to understand the relationship between fair representation learning and the promotion of fairness in the real world) (Chouldechova and Roth 2020, 88). Alternatively, the aim could be to achieve an algorithmic decision-making system that is aware of its own stereotypes and applies self-correction measures or (from an anti-subordination and anti-exclusion perspective) takes into account that certain social groups, as a collective, face worse living conditions. However, as Westen (1990, 127) stated, people share a unique and constant meaning of the concept of equality while also having a wide range of conceptions about what equality means. Therefore, the examples described represent subjective perspectives of what fairness should look like when applied by algorithms. So, who would determine the extent to which algorithms should be fairer? And based on what standards? It can be questioned whether the application of these interventions is determining, without debate, what is fair, and if, in the future, this could impact what will be considered fairness.

Thirdly, we could be facing a 'fairness washing' situation. We define this term as the use of technical fairness interventions with the objective of enhancing the perception that companies are striving to make algorithms fairer (if possible) in order to justify their use. Law (2026, 5) affirms that current 'governance frameworks reveal a dominant approach that focuses on technical compliance rather than substantive fairness'.[32] In this regard, Hartmann et al. (2026, 9-10), in an article presenting a tool for query-efficient auditing of black-box Large Language Models that exhibit systematic biases across demographic groups, alert readers that their research is a measurement

[32] The author refers to 'bias washing', describing it as a 'phenomenon where organisations meet regulatory requirements without achieving meaningful equity'.

tool, not a solution to algorithmic harm. The authors inform readers about the risks of 'ethics washing' through the misuse of their tool to give a false sense of accountability without real systemic change. To erase any suspicion of fairness washing, it would be wise to recognise the limitations of any attempt to make an algorithm fairer.

In 2020, Newman, Fast and Harmon (2020, 162) questioned whether the compilers of analytics and writers of algorithms would adapt their methods to comport with intuitive notions of morality, or if the understanding of fairness would be adjusted to changing times. Globally, both options occur simultaneously, and consequently, all the possibilities presumed for the aim of technical fairness interventions can also transpire concurrently. It is conceivable that those developing technical fairness interventions focus on the 'appearance' of fairness in AI systems without considering the broader implications: what happens when an accurate algorithm perpetually reproduces the perceptions and data correlations of an unfair society? This does not imply a lack of desire for a more egalitarian society, but their conception of how discrimination operates appears very limited, such as affirming that technological interventions alone will transform a profoundly unequal and unfair society. Conversely, the assumption that fairness is inherently complex and multifaceted, and that AI systems (and their developers) tend to oversimplify reality, is difficult to reconcile within an efficiency-based system. This perspective also precludes the proliferation or adjustment of technical methods to intuitive notions of morality that cannot be translated into technical interventions. As Srivastava et al., (2019, 2466) point out, 'existing mathematical notions of fairness are in comparison very simplistic, and they can never reflect all the nuances involved'.

VI. CONCLUSIONS

Even though we should not romanticise a non-digital past (Chander 2017, 1045), we should acknowledge that technologies have increased the complexity of addressing certain existing social issues such as discrimination.

Building on our analysis of EU law regarding fairness in AI systems, technical fairness interventions can be lawful considering the broad governance structure presented by the AI Act, which at its core includes the protection of fundamental rights. Approaching the question from an EU anti-discrimination law perspective, technical fairness interventions can be considered affirmative action measures. However, the ECJ case law on affirmative action measures restricts their implementation to a reconciliation (as far as possible) with the principle of equal treatment, thereby reducing their power to alter the more disadvantageous situation of discriminated social groups. To circumvent the requirements of ECJ case law on affirmative action, Weerts et al., (2024) have

Estudios de Deusto. Revista de Derecho Público
© Universidad de Deusto • ISSN 0423-4847 • ISSN-e 2386-9062
https://doi.org/10.18543/ed.3748 • http://www.revista-estudios.deusto.es/

proposed that these technical fairness interventions be considered a positive obligation to actively 'do no harm'. Nevertheless (and despite the interesting perspective of this proposal) its argument is based on the premise that algorithmic decision-making is neither neutral nor objective, yet human decision-making is also neither neutral nor objective. This raises the question of why different rules should apply to the former.

Beyond the discussion regarding the lawfulness of these measures, other connected questions need to be addressed, principally: what exactly is the goal of these technical fairness interventions? Is their objective to create a digital representation of a discriminatory society, to make algorithmic systems less discriminatory compared to current social discrimination, or to enhance the perception that companies are striving to make algorithms fairer, with 'fairness washing'?

The query we wish to pose concerns the fact that technical fairness interventions cannot be separated from contested legal and social understandings and debates regarding the meanings of equality. From our perspective, equality, non-discrimination and fairness must be understood in their sociological conceptualisation,[33] not as achievable goals but as social tendencies. As a society, we must tend towards equality, non-discrimination and fairness. However, even if this conceptualisation solves the issue of crystallisation (in other words, taking society as it is at a precise moment and not as an evolving system), it does not provide an answer to the reality of the multiple perceptions of equality, non-discrimination and fairness that are continuously subject to ethical, legal, doctrinal and jurisprudential debate. A non-digital society has coexisted with these ongoing debates, but is this possible in a technological society? Technical fairness interventions serve as an interesting starting point for this discussion.

VII. REFERENCES

Allhutter, Doris et al. 2020. "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective." *Frontiers of Big Data* 3(5): 1-17. https://doi.org/10.3389/fdata.2020.00005

Álvarez del Cuvillo, Antonio. 2022. "El problema de la discriminación inversa: ¿Es posible discriminar a quienes pertencen a los grupos sociales dominantes?" *Trabajo, Persona, Derecho, Mercado* 5: 187-209. https://dx.doi.org/10.12795/TPDM.2022.i5.08

Anderson, Joshua W. and Shyam Visweswaran. 2025. "Algorithmic individual fairness and healthcare: a scoping review." *JAMIA Open* 8(1): 1-8. https://doi.org/10.1093/jamiaopen/ooae149

[33] In their legal conceptualisation, equality and non-discrimination are directly applicable rights.

Aslam, Iqra. 2025 “Situating and Understanding Machine Unlearning, Ethically.” *Proceedings of Fourth European Workshop on Algorithmic Fairness* 294: 362-368.

Barocas, Solon and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review* 104: 671-732. http://dx.doi.org/10.15779/Z38BG31

Bayamlıoğlu, Emre. 2022. “The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called “right to explanation”.” *Regulation & Governance* 16(4): 1058-1078. https://doi.org/10.1111/rego.12391

Belloso Martín, Núria. 2022. “¿Se puede automatizar la fairness en la inteligencia artificial? La problemática de la discriminación y la dimensión de la interseccionalidad?”. In *Intellegentiae artificialis, imperium et civitatem,* edited by Geraldo Costa da Camino and Álvaro. A. Sánchez Bravo, 37-58. Madrid, Alma Mater.

Bower, Aamanda et al. 2017. “Fair pipelines” *arXiv*: 1-5.

Burda, Pavlo and Sieuwert Van Otterloo. 2025. “Fairness definitions explained and illustrated with examples.” *Computers and Society Research Journal* 2: 1-23. https://doi.org/10.54822/PASR6281

Calvi Alessandra and Dimitris Kotzinos. 2023. “Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective.” *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*: 1229 - 1245.

Castelnovo, Alessandro et al. 2022. “A clarification of the nuances in the fairness metrics landscape.” *Scientific Reports* 12: 1-21. https://doi.org/10.1038/s41598-022-07939-1

Caton, Simon and Christian Haas. 2024. “Fairness in Machine Learning: A Survey.” *ACM Computing Surveys* 56(7): 1-38. https://doi.org/10.1145/3616865

Chander, Anupam. 2017. “The Racist Algorithm?”, *Michigan Law Review* 115(6): 1023-1045.

Chouldechova, Alexandra and Aaron Roth. 2020. “A snapshot of the frontiers of fairness in machine learning.” *Communications of the ACM,* 63(5): 82-89.

Dwork, Cynthia et al. 2012. “Fairness through awareness.” *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf. ACM*: 214-226. https://doi.org/10.1145/2090236.209025

Dwork, Cynthia and Christina Ilvento. 2019. “Fairness under Composition.” *10th Innovations in Theoretical Computer Science Conference*: 1-20. https://doi.org/10.4230/LIPIcs.ITCS.2019.33

Fredman, Sandra. 2011. *Discrimination Law*. Oxford: Oxford University Press.

Hardt, Moritz et al. 2016. “Equality of opportunity in supervised learning.” *Advances in Neural Information Processing Systems*: 3323-3331.

Hartmann, David et al. 2026. “Audit Me If You Can: Query-Efficient Active Fairness Auditing of Black-Box LLMs.” *arXiv*: 1-26.

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

Ho, Daniel E. and Alice Xiang. 2020. “Affirmative Algorithms: The legal grounds for fairness as awareness.” *The University of Chicago Law Review Online*: 134-154.

Hock, Holly et al. 2024. “Discrimination for the sake of fairness by design and its legal framework.” *Computer Law & Security Review* 52: 1-14. https://doi.org/10.1016/j.clsr.2023.105916

Höltgen, Benedikt and Núria Oliver. 2025. "Reconsidering Fairness Through Unawareness From the Perspective of Model Multiplicity." *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*: 241-255. https://doi.org/10.1145/3757887.3763007

Holtz, Hajo M. and Jonas Ledendal. 2026. "AI Data Governance: Overlaps Between the AI Act and the GDPR." *Law, Innovation and Technology*: 1-30. https://doi.org/10.1080/17579961.2026.2633677

Kaminski, Margot E. 2019. "The Right to Explanation, Explained." *Berkeley Technology Law Journal* 34: 189-218. https://doi.org/10.15779/Z38TD9N83H

Kearns, Michael et al. 2019. "Average Individual Fairness: Algorithms, Generalization and Experiments." *33rd Conference on Neural Information Processing Systems*: 1-10.

Kleinberg, Jon et al. 2017. "Inherent trade-offs in the fair determination of risk scores." *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*: 1-23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Kleinberg, Jon et al. 2018. "Discrimination in the age of algorithms." *Journal of Legal Analysis* 10: 113-174. https://doi.org/10.1093/jla/laz001

Kroll, Joshua et al. 2017. "Accountable algorithms." *University of Pennsylvania Law Review* 165 (3): 633-705.

Law, Jason. 2026. "The ethical imperative of algorithmic fairness in AI-enabled hiring: a critical analysis of bias, accountability, and justice." *AI and Ethics* 6: 1-14. https://doi.org/10.1007/s43681-025-00927-x

Mehrabi, Ninareh et al. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54(6): 1-35. https://doi.org/10.1145/3457607

McLoughney, Aidan James et al. 2023. "'Emerging proxies' in information-rich machine learning: a threat to fairness?" *IEEE International Symposium on Ethics in Engineering, Science, and Technology*: 1.

Metikoš, Ljubiša and Jef Ausloos. 2025. "The right to an explanation in practice: insights from case law for the GDPR and the AI Act." *Law, Innovation and Technology,* 17(1): 205-240. https://doi.org/10.1080/17579961.2025.2469349

Nabi, Razieh and Ilya Shpitser. 2018. "Fair inference on outcomes." *Proceedings of the 32nd AAAI Conference on Artificial Intelligence,* 32(1): 1932-1940. https://doi.org/10.1609/aaai.v32i1.11553

Newman, David T. et al. 2020. "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions." *Organizational Behavior and Human Decision Processes,* 160: 149-167.

Nowak, Tobias. 2022. "Dutch positive action measures in higher education in the light of EU law." *Maastricht Journal of European and Comparative Law*, 29(4): 468-482.

Pessach, Dana and Erez Shmueli. 2022. "A Review on Fairness in Machine Learning." *ACM Computing Surveys,* 55(3): 1-44. https://doi.org/10.1145/3494672

Petersen, Felix et al. 2021. "Post-processing for individual fairness." *35th Conference on Neural Information Processing Systems*: 1-12.

Small, Edward et al. 2024. "Equalised Odds is not Equal Individual Odds: Post-processing for Group and Individual Fairness." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*: 1559-1578. https://doi.org/10.1145/3630106.3658989

Srivastava, Megha et al. 2019. "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning." *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*: 2459-2468. https://doi.org/10.1145/3292500.3330664

Van Bekkum, Marvin. 2025. "Using sensitive data to de-bias AI systems: Article 10(5) of the EU AI act" *Computer Law & Security Review* 56: 1-11. https://doi.org/10.1016/j.clsr.2025.106115

Van Bekkum, Marvin and Zuiderveen Borgesius, Frederik. 2023. "Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?" *Computer Law & Security Review* 48: 1-12. https://doi.org/10.1016/j.clsr.2022.105770

Wachter, Sandra et al. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7: 76-99.

Waddington, Lisa and Mark Bell. 2011. "Exploring the Boundaries of Positive Action under EU Law: A search for conceptual clarity." *Common Market Law Review,* 48(3): 1503-1526.

Weerts, Hilde et al. 2024. "The Neutrality Fallacy: When Algorithmic Fairness Interventions are (Not) Positive Action." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*: 2060-2070. https://doi.org/10.1145/3630106.3659025

Westen, Peter. 1990. *Speaking of Equality: an analysis of the rhetorical force of equality in moral and legal discourse.* Princeton: Princeton University Press.

Yao, Wei et al. 2024. "Understanding fairness surrogate functions in algorithmic fairness." *Transactions on Machine Learning Research*: 1-49.

Zhang, Cindy et al. "Matrix estimation for individual fairness." *International Conference on Machine Learning*: 40871-40887.

Žliobaitė, Indrė and Bart Custers. 2016. "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models". *Artificial Intelligence and Law,* 24(2): 183-201. https://doi.org/10.1007/s10506-016-9182-5