Equality insights in the development of fairer high-risk AI systems and the control of its discriminatory impacts

ANNA CAPELLÀ I RICART, Universitat Autònoma de Barcelona, Spain

In this paper we take as a reference the AI Act and the EU Directives on standards for equality bodies (2024/1499 and 2024/1500) with the aim to analyse how institutions can play a role in developing fairer AI systems. In parallel, we study the relevance of equality and non-discrimination experts to convey the scope and complexity of some concepts that are used in the non-discrimination field (such as intersectionality or structural discrimination) to the AI discipline, because they are not always easily traduced. We examine these questions regarding certain provisions of the AI Act that involve data governance, redress measures, the development of AI systems, the assessment of the impact on fundamental rights and the investigation regarding discriminatory results of AI systems. Furthermore, we argue that algorithmic discrimination, by shedding new light on the complex, varied and interconnected mechanisms by which discrimination operates, is pressing non-discrimination law to evolve from a simpler structure to a more sophisticated approach to inequality.

Keywords: equality, artificial intelligence, algorithms, fairness, high-risk

Reference Format:
Anna Capellà i Ricart. 2025. Equality insights in the development of fairer high-risk AI systems and the control of its discriminatory impacts. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF '25)*. Proceedings of Machine Learning Research, 16 pages.

1 Introduction

Decision-making made through AI systems can lead to discriminatory outcomes [30, 34, 35]. Situations where a face recognition software does not recognize the face of a black woman, an ad delivery is gender stereotyped, or the cost of car insurance is influenced by nationality [38] are not unusual. Furthermore, it has been highlighted that algorithmic discrimination causes less moral outrage than human discrimination [2], which can have implications in the will to challenge (or not) this discrimination, and it does not heighten guilt perception resulting from making discriminatory decisions [20], which can have implications regarding the responsibility of the discriminatory outcome.

The Regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (hereinafter, AI Act or the Regulation) is aimed at, among other objectives, ensuring a high level of protection of fundamental rights enshrined in the European Union Charter of Fundamental Rights (hereinafter, EUCFR), against the harmful effects of artificial intelligence systems (AI systems). The AI Act is structured on the basis of a risk-based approach according to the risks AI systems can generate to health, security and fundamental rights. This approach divides the uses of AI systems between unacceptable risk, high-risk, risk that implies transparency obligations, minimum risk and systemic risk.

Authors' Contact Information: Anna Capellà i Ricart, Universitat Autònoma de Barcelona, Barcelona, Spain, anna.capella@uab.cat.

EWAF '25, June 30–July 02, 2025, Eindhoven, NL

The Regulation is applied to providers and deployers of AI systems established or located within the European Union (EU) or in a third country, where the output produced by the AI system is used in the Union (Article 2.1). Different types of obligations are established mainly for the providers and deployers of AI systems and failure to comply with the provisions of the Regulation implies penalties up to 35 000 000 EUR (Article 90). High-risk systems are at the centre of the Regulation, given that their providers will have to comply with certain requirements to obtain the "CE" marking and ensure their free movement in the internal market.

The AI Act recognises the risk of AI systems to have discriminatory outcomes or to perpetuate historical patterns of discrimination (Recital 56, 57, 58, 67). In this regard, in this paper we want to analyse how perspectives from the field of equality can contribute to develop fairer AI systems and control its discriminatory impacts. These perspectives can emerge both from the doctrine and from the institutional bodies specialised in equality that in their work take into account relevant aspects for the analysis of discrimination. For instance, the possibility that a person is discriminated against at the same time on different grounds that interact and together produce a specific disadvantage [3] (intersectional discrimination) or the fact that discrimination is pervasively and deeply embedded in systems, laws, written or unwritten policies, and entrenched practices and beliefs that produce, condone, and perpetuate widespread unfair treatment [36] (structural discrimination).

Concerning the institutional bodies, two directives were approved on May 2024 on standards for equality bodies: Directive (EU) 2024/1499 on standards for equality bodies in the field of equal treatment between persons irrespective of their racial or ethnic origin, equal treatment in matters of employment and occupation between persons irrespective of their religion or belief, disability, age or sexual orientation, equal treatment between women and men in matters of social security and in the access to and supply of goods and services, and Directive (EU) 2024/1500 on standards for equality bodies in the field of equal treatment and equal opportunities between women and men in matters of employment and occupation. These Directives must be transposed by the Member States by the 19 June 2026.

These Directives have the purpose to lay down minimum requirements for the functioning of bodies for the promotion of equal treatment to improve their effectiveness and guarantee their independence in order to strengthen the application of the principle of equal treatment. They cover the mandate, independence, resources, tasks and powers of equality bodies to engage in the prevention of discrimination and awareness raising activities, to deal with cases of discrimination and to assist victims. The adoption of these directives is relevant because they exemplify the aim of the EU legislation to strengthen the role of equality bodies to combat discrimination.

Both Directive (EU) 2024/1499 and Directive (EU) 2024/1500 (together we will refer to them as “Equality bodies’ Directives” or “EBD”) have the same articles (26) that have the same content and the objective to lay down minimum requirements for the functioning of equality bodies (in the field of equal treatment between persons irrespective of their sex, racial or ethnic origin, religion or belief, disability, age, sexual orientation) to improve their effectiveness and guarantee their independence in order to strengthen the application of the principle of equal treatment (Article 1 and 3 EBD). This means that Member States may introduce or maintain provisions which are more favourable than the minimum requirements set up in the directives (Article 20 EBD). The scoop of application of the directives is limited to the grounds covered by the non-discrimination directives 79/7/EEC of 19 December 1978 on the progressive implementation of the principle of equal treatment for men and women in matters of social security; 2000/43/EC of 29

June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; 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; 2006/53/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); and 2010/41/UE of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity in a self-employed capacity.

Only in recital 22 of Directive (UE) 2024/1499 and 21 of Directive (UE) 2024/1500 there is a specific mention to automated decision making: after pointing out the opportunities of using automated systems, including artificial intelligence, it is highlighted that resources must be established to assess if automated systems are compliant with non-discrimination rules.

From a methodological perspective, to address how an approach from the field of equality can contribute to develop fairer AI systems and control its discriminatory impacts we have analysed different EU norms, namely, the AI Act, the Equality Bodies' Directives, and other EU Member States's law to provide a comparative perspective and enrich the discussion. Additionally, we have explored certain relevant legal cases, equality bodies' decisions regarding algorithmic discrimination and specialised literature to gain a more comprehensive perspective.

The paper has the following structure. First, we analyse equality and non-discrimination expertise in data governance as envisaged in the AI Act (Section 2). Next, we study the different redress measures for discriminated people in the AI Act and the Equality Bodies' Directives (Section 3). Then, we explain the incidence equality expertise can have in the development of AI systems, in the assessment of their impact on fundamental rights, and in the investigation of discrimination in AI systems (Section 4). Finally, we provide some conclusions (Section 5).

2 Equality and data governance

In this Section we will highlight how equality expertise can emphasize the wide different types of bias that can be found in the data and the implications of the lack of equality data to comply with article 10 AI Act related to data and data governance.

Equality data can be defined as any piece of information that is useful for the purposes of describing and analyzing the state of equality, and can be obtained from different sources such as, among others, official statistics, complaints data, research, diversity monitoring, population census, household surveys or administrative registers [12]. Equality data can have an impact on algorithms' quality as its precision depends on the availability of non-biased and accurate data. Furthermore, equality data can help the better identification of risks of discrimination in relation to specific groups and the proof of discrimination [27].

The adequacy model presented by the AI Act for high-risk AI systems takes into account data quality (Article 10). In the regulation is pointed out that if the AI system is not trained with high-quality data, does not meet adequate requirements in terms of its performance, its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, it may single out people in a discriminatory or otherwise incorrect or unjust manner (Recital 59).

Article 10.2 states that training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system, for example, the examination in view of possible biases that are likely to (…) have a negative impact on fundamental rights or lead to discrimination prohibited under Union law (f).

In the third paragraph, it is specified that training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. The sentence "to the best extent possible" is included because, as Northcutt et al. [26] suggest, datasets are usually not free of errors and complete. Furthermore, it is relevant to highlight that the requirements should only be met sufficiently and in view of the purpose of the system [37]. The article also specifies that datasets shall have the appropriate statistical properties, including where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used. In the fourth paragraph of the article, it is described that data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioral or functional setting within which the high-risk AI system is intended to be used. This reference to the geographical, contextual, behavioral or functional setting responds to the critics that have been made concerning the decontextualization of data. In this regard Crawford [5] references as an example the Penn Treebank Project, that gathers in the one dataset data from different sources such as stories from the Federal News Service about terrorist activities in South America, Department of Energy abstracts, Dow Jones Newswire stories or Department of Agriculture materials among others [25]. This issue persists after the approval of the AI Act and beyond the non-discrimination field as data decontextualization is classified as one of the limitations of big data for management and evaluation [39].

Finally, in the fifth paragraph of the article it is established that to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems, the providers of these systems may exceptionally process special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons. The article provides also a list of conditions that shall apply in order for such processing to occur, including the verification that only authorized persons with appropriate confidentiality obligations have access to those personal data (c), the deletion of the special categories of personal data once the bias has been corrected or the personal data has reached the end of its retention period (e), or the documentation of the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data (f).

In cases where the provider has opted for the processing of special categories of data with the purpose of ensuring bias detection and correction, equality bodies guidance can be helpful to understand the complexity of bias and recognize to what extent different types of bias can exist in automated systems. We can define bias as a disproportionate preference for (or, *a contrario*, an aversion to) an idea or a group of people originated by the following two-steps process: 1) notice that specific kinds of characteristics, for example, race, gender, age or sexual orientation, are associated with having more or less status, power, resources, or respect from others, 2) judge people or groups according to the social value associated to particular characteristics. Then, there are different types of bias that can exist, such as, gender bias, age bias, beauty bias, authority bias, name bias, sexual orientation bias, height bias, etc. Equality bodies, experts in these matters can contribute to gain a qualitative approach to AI systems' bias explaining and guiding the

process of bias detection and correction not only to most explicit bias but to other bias that can go unnoticed or promote investigations on how bias can traduce to automated bias [8].

Anyway, equality data stands out as a key element regarding non-discrimination in AI systems. However, Huddleston [21] points out that on most grounds, available equality data is usually not a systematic or obligatory part of policymaking, planning, implementation and evaluation, few national and local equality monitors have been created on equality grounds in order to raise awareness and use of equality data and equality data is rarely used in discrimination cases and the design of remedies, sanctions and positive actions. This affects directly the effectiveness of any policy regarding equality and the development of fairer AI systems.

3 Different redress measures for affected people

In this Section we analyze different redress measures for people who consider that they have suffered discrimination envisaged in the AI Act and the Equality Bodies' Directives.

The AI Act recognizes in Article 86 the right to receive an explanation of individual decision-making (that won't be studied further in this paper) and the right to lodge a complaint with a market surveillance authority in the case of an infringement of a provision of the regulation (Article 85).

Complaints with a market surveillance authority, according to article 11 of Regulation (EU) 2019/1020 on market surveillance and compliance of products, are a factor in deciding on which checks to perform, on which types of products and on what scale, and market surveillance authorities are obliged to establish procedures for following up on complaints or reports on issues relating to risks or non-compliance. In consequence, if someone that has been discriminated against with an AI system lodges a complaint in virtue of article 85 AI Act, the complaint will have the purpose to notify a violation of the Regulation and not to receive assistance from the market surveillance authority. Indeed, the market surveillance authority does not seem to be allowed to act on behalf of the complainant -because as a market surveillance authority it's not its goal.

On the contrary, equality bodies will be able to provide assistance to victims, receive complaint of discrimination (that according to article 12 EBD can be orally, in writing and online and with no costs) and to inform the complainants of the legal framework or the available remedies (among other items) and whether the complaint will be closed or there are grounds to pursue it further (Article 6 EBD). Moreover, they will be able to issue binding opinions or non-binding decisions and to offer the parties the possibility of seeking an alternative resolution to their dispute -such as, mediation or conciliation, in accordance with national law and practice- (Article 7 EBD).

Another function that the Equality bodies' Directives entrust to these organisms is the right to act in court proceedings in civil and administrative law matters (Article 10 EBD). This right shall include, at least one of the following: the right to initiate court proceedings on behalf of one or several victims; to participate in court proceedings in support of one or several victims; to initiate court proceedings in its own name, in order to defend the public interest; or to act as a party in proceedings on the enforcement or judicial review of a binding decision issued by the equality body. For example, the Spanish Independent Authority of Equal Treatment and non-discrimination can initiate court proceedings on behalf of one or several victims and initiate court proceedings in its own name, in order to defend the public interest (Article 11.bis of Spanish Civil Procedure Law). In France, in contrast, the Defender of rights can

exclusively participate in court proceedings in support of one or several victims (Article 33 de la LOI organique n° 2011-333 du 29 mars 2011 relative au Défenseur des droits). In the Commission Staff Working Document, "Equality bodies and the implementation of the Commission Recommendation on standards for equality bodies", it is highlighted that equality bodies' power to litigate in their own name, implying that they can engage in cases of collective discrimination to represent the public interest without there being a specific complainant, can be particularly beneficial in tackling cases of discrimination caused by artificial intelligence [10]. This is because it has been highlighted that individual litigation in the field of algorithmic discrimination can be very difficult because it is difficult to gather evidence [23, 9, 13]. Regarding cases of collective discrimination, it is pertinent to mention that article 110 AI Act includes this Regulation in the scope of application of Directive (EU) 2020/1828 on representative actions for the protection of the collective interests of consumers.

From our perspective, private litigation should be complemented by powerful mechanisms of public enforcement, that with adequate resources could be executed by equality bodies. However, equality bodies do not have the government support (which translates, among other aspects, into the provision of economic resources) that other supervision agencies do enjoy. For example, in Spain, the recently created Artificial Intelligence's Supervision Agency (AESIA) had an allocation of 5 million euros for its implementation in the 2022 General Budgets (Law 22/2021, of 28th December, of General Budgets of the State for the year 2022). In contrast, the Spanish Independent Authority for the equal-treatment and non-discrimination envisaged by the Law 15/2022, of 12th of July, for the equal treatment and non-discrimination, has not been yet created, despite the fact that its constitution was presupposed within six months of the law's entry into force (First Addition provision of the Law 15/2022).

4 Equality expertise in development, impact assessment of fundamental rights, and investigation of AI systems

In this Section we study how equality expertise can impact in the development of AI systems (4.1), the performance of the assessments of the impact on fundamental rights (4.2) and the investigation of the discriminatory impacts that AI systems can have (4.3).

4.1 Development of AI systems

Concerning the advice on the development of AI systems, on the one hand, the AI Act includes an advisory forum with a view to ensuring the involvement of stakeholders in the implementation and application of the Regulation (Recital 150) and to provide technical expertise and advice the Board and the Commission (Article 67). Moreover, the European Commission must establish an independent scientific panel of experts intended to support the enforcement activities (Article 68 AI Act). Even though the composition of this panel is not published yet, it should include experts on equality and non-discrimination. In this regard, in article 3.3.c of the Draft for setting out rules for the establishment and operation of a scientific panel of independent experts in the field of artificial intelligence, published by the European Commission to receive feedback from stakeholders, it is established that selected experts must ensure multidisciplinary adequate and up-to-date scientific or technical expertise in the field of artificial intelligence, including (…) fundamental

rights and equality (…). On the other hand, Equality bodies' Directives point out that there must be procedures that ensure that government and relevant public authorities can consult equality bodies on legislation, policy, procedure and programs related to the rights and obligations derived from antidiscrimination directives (Article 15).

Furthermore, the AI Act establishes the creation of AI regulatory sandboxes to provide for a controlled environment that fosters innovation and facilitates the development, training, testing and validation of innovative AI systems for a limited time before being placed on the market or put into service (Article 57.4).

Competent authorities shall provide, as appropriate, guidance, supervision and support within the AI regulatory sandbox with a view to identifying risks, in particular to fundamental rights (…), testing, mitigation measures, and their effectiveness in relation to the obligations and requirements of the Regulation (Article 57.5). Furthermore, national competent authorities shall ensure that, to the extent the innovative AI systems involve the processing of personal data or otherwise fall under the supervisory remit of other national authorities or competent authorities providing or supporting access to data, those authorities are associated with the operation of the AI regulatory sandbox and involved in the supervision of those aspects to the extent of their respective tasks and powers (Article 57.10). Conceivably, equality bodies may be involved in the operation of controlled testing spaces. Any significant risks to the fundamental right of non-discrimination identified during the development and testing of such AI systems shall result in adequate mitigation or the temporary or permanent suspension of the testing process, or the participation in the sandbox if no effective mitigation is possible (Article 57.11).

AI sandboxes provide an opportunity to identify the system's vulnerabilities and the potential effects it may have on people or in the economic, social or ecological sphere and allow authorities to encourage the adoption of ethical practices for the benefit of society [14]. The point of view of equality experts and equality bodies can be relevant to analyzing AI systems outputs results from a perspective that takes into account specific concepts of the non-discrimination doctrine, such as intersectionality or structural discrimination. Additionally, equality experts and equality bodies must cooperate with experts from other disciplines to address equality concerns. As an example, concerns around fairness relate to algorithmic bias and discrimination, the use of personal data (including sensitive personal data) and sector-specific notions of fairness such as financial fairness [29]. Besides, AI systems tested in the sandboxes can have an incidence guiding equality bodies' work program (Article 17 EBD) to investigate and analyze the discriminatory outputs found. In any case, further research must be done regarding the introduction of equality expertise in AI sandboxes, providing clear parameters of how to look for the system's vulnerabilities and potential effects concerning equality and non-discrimination.

Furthermore, equality bodies must also publish reports with recommendations, at least every four years, on the state of equal treatment and discrimination, including potential structural issues, in the member state (Article 17 EBD). AI systems can have an incidence on structural issues in Member States, taking into account that they can participate in the perpetuation of certain stereotypes or forms of exclusion and these questions should be considered in equality bodies' reports on the state of equal treatment and discrimination. For example, equality bodies should monitor if visual ageism, understood as the representation of older people being in peripheral or minor roles without positive attitudes [22], can be found in images available in the Member State used by AI systems, or if there are enterprises in the Member State that utilize hiring algorithms that use as a relevant factor the date in which a candidate graduated from school [33], to

prevent people from certain age to participate in a recruitment process. Research done in these fields could then have an impact on the development of AI systems.

It is pertinent to point out that each Member State must designate a national competent authority that has the task of providing guidance on the implementation of the AI Act (Article 70 AI Act). It is specified that national competent authorities shall have enough personnel permanently available whose competences and expertise shall include an in-depth understanding of AI technologies, data and data computing, personal data protection, cybersecurity, fundamental rights, health and safety risks and knowledge of existing standards and legal requirements. Conversely, in the Equality Bodies' Directives it is only specified that equality bodies are provided with the human, technical and financial resources necessary to perform all their tasks and to exercise all their competences effectively (Article 4). It would have been interesting that among the personnel envisaged to work in equality bodies by Equality Bodies' Directives, professionals on AI would have been included.

4.2 Guidance regarding the assessment of the impact on fundamental rights, taking into account intersectionality and structural discrimination

According to Article 27 AI Act, before placing on the market an AI system of Annex III, an assessment of the impact on fundamental rights (AIFR) shall be performed by deployers that are bodies governed by public law, deployers that are private entities providing public services, and all deployers of AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score and AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance.

The AIFR must include the categories of natural persons and groups likely to be affected by the use of the AI system in the specific context (Article 27.1.c), the specific risks of harm likely to have an impact on the categories of persons or groups of persons identified pursuant point c (Article 27.1.d), and the measures to be taken where those risks materialize (Article 27.1.f). A summary of the findings of the AIFR conducted (Section C of Annex VIII) shall be provided and thereafter kept up to date by the deployer to be included in the EU database for high-risk AI systems listed in Annex III.

From an equality perspective, the AIFR presents some shortcomings. Firstly, the obligation applies to the first use of the high-risk AI system (Article 27.2), in consequence, it is not periodic or mandatory through all the AI system lifecycle. This assessment should be repeated through all the system's lifecycle and the results (not only a summary) should be public to increment transparency and promote the review by civil society. Also, there can be several fundamental rights affected, and not only for the use of a particular AI system, but for the systemic use of AI. It has been analyzed, in this regard, for example, how AI systems can affect critical thinking [19], key to a fundamental right such as freedom of thought (Article 10 EUCFR) but also a pilar of the achievement, protection and development of any other fundamental right.

These considerations can have an impact in the performance and the evaluation of the output of assessments on the impact on fundamental rights, that cannot be simple but adapted to analyze the interaction of the high-risk AI system

on a complex society. The AI Act does not do any reference to intersectional discrimination, so equality expertise can be useful in this matter.

For example, Alderman et al. [1] have proposed the STANDING Together recommendations, a set of suggestions with the aim to encourage transparency regarding limitations of health datasets and proactive evaluation of their effect across population groups. Recommendation 2.2e specifies the need to identify disparate performance in any additional groups outside of the prespecified contextualized groups of interest that can be identified manually considering a longlist of attributes, including (but not limited to) sex, gender, race, ethnicity, age, socioeconomic status, sexual orientation, disability, pregnancy, relationship or marriage status, religion or belief, nationality, ancestry, occupation, language or languages spoken, caste, creed, or tribe. It is relevant that in the STANDING Together recommendations intersectionality is defined as how individuals are subject to several, intersecting forms of privilege or disadvantage. In this regard, it is the competence of experts and equality bodies to identify all factors that can lead to privilege or disadvantage, and this expertise must be taken into account in the evaluation of fairness in algorithmic systems.

Particularly, it would be interesting to further investigate whether a segregated normative approach of the discriminatory grounds, followed by the European union and members states, is the best option to address algorithmic discrimination. In depth knowledge of how discrimination in the data, source codes, algorithms, and software occur can reveal that the combination of all these tools results in most cases in intersecting discriminations and not in differentiated (by gender, sex, age, sexual orientation, disability or economic resources) discriminations, as it is approached by the current normative.

It must be acknowledged, although, that academic works that intend to tackle intersectional discrimination in AI systems [24, 28] go beyond what European non-discrimination law actually obliges to. Nor do the Treaties of the EU, the EUCFR or the EU anti-discrimination directives establish obligations regarding intersectional discrimination. This is interesting because a more pressing reality - the fact that patently AI systems discriminate taking into account different personal characteristics- is pressing non-discrimination law to evolve and to understand that most of the time there is an interaction between different discriminatory motives. In this regard, maybe AI systems have shed some light or made more evident to a wider public (even though in specialized literature it has been acknowledged for a long time [6]) a deeper understanding of the complex, varied and interconnected mechanisms by which discrimination operates.

The assessment of the impact on fundamental rights should also take into account if the system promotes structural discrimination even though it can be an abstract concept. The AI Act refers to it through expressions such as “the exclusion of certain groups” (Recital 31), “perpetuate historical patterns of discrimination” (Recital 56, 57, 58), or “perpetuate and amplify existing discrimination” (Recital 67). In the AIFR the potential of the AI system to be a mechanism for the perpetuation of structural discrimination should be considered, and also the measures implemented to mitigate this risk. The AIFR should examine if mitigation measures have been implemented by the provider such as the test of the AI system to control if it has discriminatory outputs, the adaptation of methodologies of testing bias in everyday life, including tests regarding job applications, where the applicant’s name is changed to (indirectly) identify ethnicity, or advanced statistical analysis to detect differences in datasets concerning protected groups [11].

The methods to assess the discriminatory impacts of AI should take into account reports, recommendations, independent surveys or statistics published by equality bodies (Article 16 EBD), that should rely on equality data.

4.3 Advice on how to investigate algorithmic discrimination

Article 74.12 AI Act concretes that where relevant and limited to what is necessary to fulfil their tasks, the market surveillance authorities shall be granted full access by providers to the documentation as well as the training, validation and testing data sets used for the development of high-risk AI systems. Also, Article 74.10 AI Act states that these authorities shall be granted access to the source code of the high-risk AI system, if access is necessary to assess the conformity of a high-risk AI system or testing or auditing procedures and verifications based on the data and documentation provided by the provider have been exhausted or proved insufficient. In the non-discrimination field access to the source code has been examined as a way to obtain information from the algorithm regarding its discriminatory output, even though some authors argue that most of the time the source code cannot give relevant information [4, 32]. It must be considered the architecture of the algorithm or the AI system, whether it is a simple algorithm or a predictive algorithm, more complex and where the training data becomes more important.

In any case, Article 74.14 AI Act stablishes that any information or documentation obtained by market surveillance authorities shall be treated in accordance with the confidentiality obligations set out in Article 78, meaning that most information gathered by the market surveillance authority has confidentiality obligations that diminish its potential subsequent use with the purpose to discourage offenders to repeat its discriminatory actions (for the dissuasive effect in algorithmic discrimination cases see: Judgment 2949/2020, of 31 of December, of the Bolognia Court). In fact, in the AI Act there are not envisaged any dissuasive mechanisms, which in the non-discrimination field are relevant. In this regard, it is interesting to point out as that the proposed Canadian Artificial Intelligence and Data Act, introduced as part of Bill C-27, established the possibility for the Minister to publish, with the purpose of encouraging compliance, information that relates to an artificial intelligence system when there is a risk of imminent harm (Article 28.1) or about any contravention of the Act on a publicly available website (Article 27.1).

Meanwhile, equality bodies will be empowered to conduct an inquiry into whether a breach of the principle of equal treatment has occurred (Article 8.1 EBD). It is specified that member states must provide a framework for conducting inquiries to carry out fact-finding with effective rights to access information and documents that are necessary to establish whether discrimination has occurred (Article 8.2 EBD). For example, public administrations and individuals must provide the necessary collaboration to the Spanish Independent Authority of Equal Treatment and non-discrimination providing all kinds of information and data that they have in its possession (Article 44 of the Spanish Law 15/2022, of 12th of July, for the equal treatment and non-discrimination). The French Defender of rights can request explanations and information, request documents, make audits and verifications on site [16]. The Deutch Independent Federal Commissioner for Anti-Discrimination is able to require information to all the federal ministries, other federal authorities and federal bodies governed by public law (Section 28 of the Deuch General Act on Equal Treatment). The prerogative to make inquiries can be very useful to gain information on whether the algorithmic system has a discriminatory impact.

Equality bodies must be able to provide and document their assessment of the case, including establishing the facts and a reasoned conclusion on the existence of discrimination (Article 9 EBD). The article gives the option that this

assessment takes form of non-binding opinions or binding decisions. For example, the Franch Defender of rights, issued, on 8th April 2019, the Decision nº 2019/099, related to the question of a student union to the president of the University of Y to transmit the algorithmic processes used by the commissions responsible to examine the license entry applications submitted on Parcoursup as well as the corresponding source codes.

Both opinions and decisions shall include specific measures to remedy any breach of the principle of equal treatment found and to prevent further occurrences and establish the appropriate mechanisms for the follow up to non-binding opinions and the enforcement of binding decisions. Equality bodies must publish, at least, a summary of those opinions or decisions that consider especially important. The French Defender of Rights, since 2011 -when he acquired the competence regarding issues in the field of non-discrimination, previously competence of the Haute autorité pour lutter contre les discriminations et pour l'égalité (HALDE), has issued more than 1300 non-binding opinions.

Market surveillance authorities in the AI Act have access to information regarding high-risk AI systems but equality bodies, according to the Directives, do not have restrictions when it comes to whether a breach of the principle of equal treatment has occurred, so member states' equality organisms can use its prerogatives with any AI system whether or not is high risk. The different approaches and functions of the authorities oblige the existence of cooperation mechanisms between them.

For that matter, both in the AI Act and the Equality Bodies' Directives it is specified that the authorities must have cooperation mechanisms. In article 74.10 AI Act it is established that Member States shall facilitate coordination between market surveillance authorities and other relevant national authorities or bodies which supervise the application of legislation that might be relevant for the high-risk AI systems referred to in Annex III. Likewise, in Article 14 of the Equality Bodies' Directives it is specified that equality bodies must have mechanisms to cooperate with other equality bodies within the Member State, relevant public and private entities, including labor inspectorates, the social partners, and civil society organizations, at national, regional and local level as well as in other Member States and at Union and international level. It would have been interesting if the article had made specific reference to Data Protection Authorities and AI supervision authorities, taking into account that are the ones with whom equality bodies must interact to have a full picture of how different types of discrimination emerge in the context of automated decisions.

In the course of the investigation two relevant sources of information can be taken into account regarding discriminatory output. On the one hand, EU database must contain (Article 71), concerning high-risk AI systems referred to in Annex III (excluding certain exceptions), information from the provider such as, name, address and contact details, AI system trade name and any additional unambiguous reference allowing the identification and traceability of the AI system, a description of the intended purpose of the AI system and of the components and functions supported through this AI system, a basic and concise description of the information used by the system (data, inputs) and its operating logic (Annex VIII, Section A). Moreover, the EU database must contain information from the deployer, such as, name, address and contact details and a summary of the findings of assessment of the impact on fundamental rights (Annex VIII, Section B). According to article 71.4 AI Act all this information will be accessible and publicly available in a user-friendly manner, so it can complement any investigation made in relation to high-risk AI systems with possibly discriminatory outcomes. On the other hand, Article 11 AI Act states that the provider shall draw up the technical documentation of a high-risk AI system before that system is placed on the market or put into service and shall be kept

up-to date. Paragraphs 2.b, 2.d, 2.g and 3 of Annex IV (that details the documentation that must be included in the technical documentation) are of special interest because they can reveal the discriminatory outcomes that the AI systems can produce. Specifically, paragraph 2.g specifies the need for a description of the potentially discriminatory impacts. This request is good to exemplify how, for example, if this description is done by someone that is not specialized in equality and non-discrimination, possible discriminatory impacts can be underestimated. For example, the description can be done only regarding the most studied motives that cause discrimination. In this regard, Stypińska [33] states that there are other motives (rather that sex and ethnic origin) that must be taken into account and that for now have been neglected in research regarding algorithmic discrimination. The author argues that, in the case of age discrimination, for example, this can be because of different factors such as relatively weak social representation of rights of older people in the area of AI, or ideologies and stereotypical beliefs about older persons as users or non-users of AI applications held by software producers.

Furthermore, we must remain vigilant to prevent that professional from technical disciplines, with the purpose to create fairer algorithmic systems, make suitable concepts of non-discrimination by simplifying and redefining them, underestimating social complexity and its contradictory core. For example, is important to understand that if, following Friedman and Nissenbaum [17], Danks & London [7] and Ferrer et al. [15], there are three types of bias in computer systems (pre-existing from social institutions, technically created and emerging from the context of use) interventions to mitigate algorithmic bias only have impact in one of the three types (technically created), but the other ones go beyond the scoop of technical intervention of algorithms and present a complexity that cannot be technologically addressed. Being aware of the limits of technical solutions helps understand why discrimination in software even when fairness is an explicit design goal is common [18].

Continuing on the investigation of discriminatory output of AI systems, any documentation created or maintained under the Regulation in relation to the use of high-risk AI systems referred to in Annex III, can be requested and accessed by national public authorities or bodies which supervise or enforce the respect of obligations under Union law protecting fundamental rights, including the right to non-discrimination, when is necessary for effectively fulfilling their mandates within the limits of their jurisdiction (Article 77.1 AI Act). In addition, Article 77.3 AI Act states that if the documentation requested is insufficient to ascertain whether an infringement of obligations under Union law protecting fundamental rights has occurred, the public authority or body may make a reasoned request to the market surveillance authority, to organize testing of the high-risk AI system through technical means.

This information will be available in the technical documentation of high risk AI systems but in the event that the equality body wants to investigate a system that is not considered high-risk this information should be also available or it would be necessary that a regulation obliges that all algorithmic systems applied to the social context (high, low or minimal risk) have available information on the discriminatory and societal impact of the AI system, because even though it is not categorized as high risk, when applied to a certain context, the algorithm can have discriminatory results.

When the investigation is over, Equality Bodies' Directives do not make any reference to the power of equality bodies to sanction if they determine that there has been a violation of the right to equality and non-discrimination. Notwithstanding that fact, national laws can grant this prerogative. It has not been the case of the Spanish or the French equality bodies, for example, considering that both must refer to another authority vested with the power to initiate

disciplinary proceedings the facts that the body considers that justify a sanction (Article 40 of the Spanish Law 15/2022, of 12th of July, for the equal treatment and non-discrimination and Article 29 of the Franch Organic Law n° 2011-333 of 29th March 2011 of the Defender of rights. Rey Martínez [31] has criticized that the Spanish Independent Authority for equal treatment and non-discrimination cannot sanction and must refer the facts to another authority vested with the power to initiate disciplinary proceedings. He argues that it would have been better to emulate Data Protection Authorities prerogatives, that are granted the power to sanction (Article 58 of the GDPR). Less powers has the Deutch equality body that does not have any capacity to impose sanctions for the breach of the principle of equal treatment. As is specified in section 27 of the Deutch General Act on Equal Treatment, the tasks of the Deutch Federal Anti-Discrimination Agency do not include imposing sanctions.

In the field of AI, it is not stipulated in the AI Act which organism will decide on the sanctions. Article 99.9 AI Act establishes that depending on the legal system of the Member States, the rules on administrative fines can be imposed by competent national courts or by other bodies, as applicable in those Member States.

5 Conclusion

In the analysis of how equality expertise can be useful to develop fairer AI systems taking into account the provisions established in the AI Act and in the Equality Bodies' Directives, we are capable of reaching some conclusions.

From an institutional perspective, it seems that even though in the field of artificial intelligence, equality and non-discrimination are issues that must be considered by the supervisory authorities (e. g. competences and expertise of the personnel of the national competent authorities shall include an in depth understanding on fundamental rights), in the field of equality and non-discrimination, risks regarding AI systems are not considered as a main matter at a institutional level (e. g. in the equality bodies directives only in recital 21 and 22 there is a reference to automated systems and the personnel of the body is not bound by the Directives to have any competences and expertise in AI).

The market surveillance authorities envisaged in the AI Act have the ability to sanction the noncompliance with the Regulation but do not have the competence to help provide redress to people who have been discriminated against. On the contrary, equality bodies can provide redress to people who have been harmed (e.g. litigation or issuing binding decisions) but are not obliged by the Equality Bodies' Directives to establish a mechanism to sanction infractions.

Also, market surveillance authorities in the AI Act have access to information regarding high-risk AI systems and equality bodies, can have access to any information regarding whether a breach of the principle of equal treatment has occurred. We understand here that the prerogative applies independently whether the AI system is high-risk or not. However, linking it with the fact that the Equality Bodies' Directives do not oblige that these organisms' personnel include technical experts in AI, it does not matter if equality bodies can access more information if they do not know which information they must look for.

In addition, we have pointed out that it would have been interesting if the AI Act had included dissuasive mechanisms, as they have been used in non-discrimination cases. Moreover, we have highlighted that equality and non-discrimination expertise is a key aspect to the best application of the AI Act requirements for detecting and correcting biases, developing fairer AI systems, and assessing and investigating their discriminatory impact. A formal, simple or limited understanding of AI systems' discriminatory output and the wider impact they can have in society (e. g. millions

of people using an AI system that has avatars that reproduce gender stereotypes, age stereotypes or normative bodies or millions of people being excluded form a good or a service because data used leads to exclusion) can have deleterious effects.

Finally, the field of equality and non-discrimination is characterized by its ongoing and evolving doctrinal discussions that are complex and difficult to adapt to AI systems such as intersectionality or structural discrimination, being this the reason why perspectives from experts in this field and equality bodies remain crucial to ensure 'fairer' AI systems. Anyhow, algorithmic discrimination is being useful to enlighten the opaque mechanisms by which discrimination operate, understanding its *modus operandi* better and deeper, and pressing non-discrimination law to evolve from a simpler structure to a more sophisticated approach to inequality.

References

[1] Joseph E. Alderman et al. 2025. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digital Health, 7, e64–88.

[2] Yochanan E. Bigman et al. 2023. Algorithmic Discrimination Causes Less Moral Outrage Than Human Discrimination. J. Exp. Psychol. Gen, 152(1), 4-27. https://doi.org/10.1037/xge0001250

[3] Emmanuelle Bribosia, Robin Médard Inghilterra and Isabelle Rorive. 2021. Discrimination intersectionnelle en droit : mode d'emploi. Revue trimestrielle des droits de l'Homme, 126, 241-274. https://doi.org/10.3917/rtdh.126.0241

[4] Céline Castets-Renard. 2019. Accountability of Algorithms in the GDPR and beyond: A European Legal Framework on Automated Decision-Making. Fordham Intellectual Property, Media and Entertainment Law Journal, 30(1), 91-137. https://ir.lawnet.fordham.edu/iplj/vol30/iss1/3

[5] Kate Crawford. 2021. *Atlas of AI*. New Haven and London, Yale University Press.

[6] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: a Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago legal forum, 1989(1), 139-167. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

[7] David Danks and Alex John London. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 4691–4697. https://doi.org/10.24963/ijcai.2017/654

[8] Deutch Federal Anti-Discrimination Agency. 2023. Coded Bias. The General Equal Treatment Act and protection against discrimination by algorithmic decision-making systems. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/EN/publikationen/ki_study.pdf?__blob=publicationFile&v=2

[9] Raluca Dimitru. 2024. Using algorithms in personnel recruitment: some implications in labour law. Legal and Administrative Studies, Supplement – Year 2024, 12-25.

[10] European Commission, Commission Staff Working Document, *Equality bodies and the implementation of the Commission Recommendation on standards for equality bodies*, Brussels, 19.3.2021 SWD (2021) 63 final.

[11] European Union Agency for fundamental Rights. 2020. *Getting the future right: Artificial intelligence and fundamental rights*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2021-artificial-intelligence-summary_en.pdf

[12] European Commission: Directorate-General for Justice and Consumers and Timo Makkonen. 2016. European handbook on equality data – 2016 revision. https://data.europa.eu/doi/10.2838/39707

[13] Franck Fagan. 2024. *Reducing Proxy Discrimination*. arXiv:2501.03946

[14] Kate Francis. 2025. The need for an ethical approach to regulatory sandboxes. In Filippo Bagni and Fabio Seferi (eds.) Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders, CINI's Cybersecurity National Lab., Rome, 192-206.

[15] Xavier Ferrer et al. 2021. Bias and discrimination in AI: A cross-disciplinary perspective. IEEE Technology and Society Magazine, 40(2), 72–80. 10.1109/MTS.2021.3056293

[16] French Defender of Rights. 2020. *Discriminations liées à l'orientation sexuelle et l'identité de genre*. DDD_depliant_discriminations-orientation-sexuelle-et-identite-genre_20200514.pdf

[17] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. In *ACM Transactions on Information Systems (TOIS)*, 14(3), 330–347. https://nissenbaum.tech.cornell.edu/papers/Bias%20in%20Computer%20Systems.pdf

[18] Sainyam Galhotra et al. 2017. Fairness Testing: Testing Software for Discrimination. In *Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering*. Paderborn, Germany, 13 pages. https://doi.org/10.1145/3106237.3106277

[19] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies, 15(1), 6, 1-28. https://doi.org/10.3390/soc15010006

[20] Maryam Ghasemaghaei and Nima Kordzadeh. 2024. Understanding how algorithmic injustice leads to making discriminatory decisions: An obedience to authority perspective. Inf. Manag, 61(2), 1-14. https://doi.org/10.1016/j.im.2024.103921

[21] Thomas Huddleston. 2017. *Methodological approach Overview per EU Member State Technical annex*, European Commission. https://www.humanconsultancy.com/downloads/689-equality-data-collection-indicators

[22] Loredana Ivan et al. 2020. Mitigating Visual Ageism in Digital Media: Designing for Dynamic Diversity to Enhance Communication Rights for Senior Citizens. Societies, 10(4), 76, 1-13. https://doi.org/10.3390/soc10040076

[23] Aislinn Kelly-Lyth. 2023. Algorithmic discrimination at work. European Labour Law Journal, 14(2), 152–171. 10.1177/20319525231167300

[24] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Seoul, Republic of Korea, 485-494. https://doi. org/10.1145/3531146.3533114

[25] Marcus Mitchell et al. 1993. Building a Large Annotated Corpus of English: The Penn Treebank, Computational Linguistics 19(2), 313-330. https://aclanthology.org/J93-2004.pdf

[26] Curtis G. Northcutt et al. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv:2103.14749

[27] Carsten Orwat. 2020. Risks of Discrimination through the Use of Algorithms. A study compiled with a grant from the Federal Anti-Discrimination Agency. https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf%3F__blob%3DpublicationFile%26v%3D1

[28] Anaelia Ovalle et al. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23).* Association for Computing Machinery, New York, 496–511. https://doi.org/10.1145/3600211.3604705

[29] Ruth Plato-Shinar and Andrew Godwin. 2025. Regulatory Cooperation in AI Sandboxes: Insights from Fintech, EBI Working Paper Series 2025 – no. 189, Frankfurt.

[30] Vera Lúcia Raposo (2023) When facial recognition does not 'recognise': erroneous identifications and resulting liabilities. AI & Soc, 39, 1857–1869. https://doi.org/10.1007/s00146-023-01634-z

[31] Fernando Rey Martínez.2023. La (jurídica) igualdad de trato (de origen norteamericano) y la (política) igualdad de oportunidades (de origen europeo): conexiones, confusiones y malentendidos. IgualdadES, 9, 13-43. https://doi.org/10.18042/cepc/IgdES.9.01

[32] Pilar Rivas Vallejo. 2021. Discriminación algorítmica: detección, prevención y tutela. In *XXXI Jornades Catalanes de Dret Social ("Treball, discriminació i COVID")*, Associació Catalana de Iuslaboralistes, Barcelona, 1- 79.

[33] Justyna Stypińska. 2021. Ageism in AI: new forms of age discrimination in the era of algorithms and artificial intelligence. In *Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI, CAIP 2021*, Bologna, Italy, 1-9.

[34] Justyna Stypińska. 2023. AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. AI & Soc, 38, 665-677. https://doi.org/10.1007/s00146-022-01553-5

[35] Nenad Tomasev et al. 2024. Manifestations of xenophobia in AI systems. AI & Soc. https://doi.org/10.1007/s00146-024-01893-4

[36] UN Human Rights Council. 2021. Promotion and Protection of the Human Rights and Fundamental Freedoms of Africans and of People of African Descent against Excessive Use of Force and Other Human Rights Violations by Law Enforcement Officers (OHCHR 2021) Report of the UN High Commissioner for Human Rights, UN Doc A/HRC/47/53. https://www.ohchr.org/en/documents/reports/ahrc4753-promotion-and-protection-human-rights-and-fundamental-freedoms-africans

[37] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act. Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 4/2021, 97-112. https://doi.org/10.9785/cri-2021-220402

[38] Jessica Wulf. 2022. *Automated Decision-Making Systems and Discrimination: Understanding Causes, Recognizing Cases, Supporting Those Affected. A guidebook for anti-discrimination counseling*. Algorithm Watch. https://algorithmwatch.org/en/autocheck/

[39] Pete York and Michael Bamberger. 2025. The Applications of Big Data to Strengthen Evaluation. In Steffen Bohni Nielsen, Francesco Mazzeo Rinaldi and Gustav Jakob Petersson (eds) Artificial Intelligence and Evaluation, Routledge, London, 37–55.