Discriminatory bias (and how to prevent it) in the European Union: a review of the ethical and regulatory framework for artificial intelligence.

Pablo Cerezo Martínez, Alejandro Nicolás Sánchez, Francisco J. Castro-Toledo[1]

Plus Ethics, Spain

Abstract

The European Union and some of its institutions have taken significant steps to address the challenges posed by the development and use of Artificial Intelligence (AI) in various contexts. The ubiquity of AI applications in everyday life, affecting both citizens and professionals, has made AI a common topic of discussion. However, alongside its progress, concerns have been raised about the potential negative consequences of AI, in particular discriminatory bias. This article aims to examine the challenges of defining, identifying and mitigating discriminatory bias in AI systems from two perspectives: 1) to conduct an ethical and normative review of European Commission documents from the last 8 years (from GDPR to AI Act proposal); and 2) to provide practical recommendations for key stakeholders, including designers, end-users and public authorities, to minimise/mitigate this risk. The documentary review has been conducted on 21 EU normative and ethical guidelines in the field of AI, noting first, that there is no clear conceptual framework on the issue at the European level, and second, that this lack of a clear conceptual framework may affect the concreteness and detail of the potential mitigation/mitigation measures proposed.

Keywords: Artificial Intelligence, Discriminatory bias, Europe, Ethics, Regulatory framework.

[1] Correspondence: pcerezo@plusethics.com; anicolas.plusethics@gmail.com; fcastro@plusethics.com

1. INTRODUCTION

The European Union (EU) and its institutions have made numerous efforts to identify and address the challenges posed by the development and implementation of artificial intelligence (AI) in the various contexts in which it is intended to be used. This has highlighted the cross-cutting nature of these tools, which can be applied in virtually all contexts of daily life, both for citizens and for professionals in different fields; it is this proliferation of applications and the gradual improvement of tools, making them more powerful and efficient, that has led to AI becoming a topic of common discussion. However, in parallel with this progress, institutional voices have become increasingly vocal about the potential negative impact of these tools, including the issue of discriminatory bias[2].

In Europe, several concrete examples of bias in artificial intelligence have recently been identified. To name just a few of the most recent, in the field of recruitment, for example, AI tools were used that turned out to be biased against women. This happened when the AI was based on CVs submitted over the last 10 years, most of which belonged to men, leading the algorithm to favour men over women (Dastin, 2018). This trend of using AI in recruitment is expected to continue in 2024, although measures are also being taken to reduce the risk of bias. Another recent example of AI bias in Europe is the scandal in the Netherlands, where the government used an algorithm to predict who was likely to fraudulently claim child benefit. Without any evidence of fraud, the tax authorities forced 26,000 parents, targeting dual nationals and ethnic minorities, to pay back tens of thousands of euros with no right of appeal. The Dutch Data Protection Authority found the tax authorities' methods to be 'discriminatory' (Heikkilä, 2022; Henley, 2021).

The aim of this article is to address the challenges of defining, identifying and minimising discriminatory bias in AI systems within a European scope (rather guarantee-based, from an international comparative perspective) from a double point of view: a) based on an ethical and normative review of the reference documents published by European public bodies (and its different working groups) over the last 8 years, and b) with an applied purpose for the main stakeholders (designers, consumers, public authorities, etc.). To achieve this, the following content structure is proposed: in the first section, the concept of algorithmic discrimination will be introduced from a multidisciplinary perspective; in the second section, the main results of the quantitative and qualitative systematic review of the approach to the issue of discriminatory bias in the main European regulatory instruments and recommendations related to the design, development, implementation and use of AI systems will be presented; and finally, a third section will aim at systematising the recommendations to minimise and mitigate this risk. In short, this proposal makes it possible

[2] The AI risks that have raised the most concern include the following: 1) AI algorithms can perpetuate and amplify existing biases in the data, leading to discriminatory outcomes (bias and discrimination) (Mayson, S, 2019); 2) many AI models, especially the more advanced ones, are 'black boxes' that provide little or no insight into how they reach their conclusions (lack of transparency) (Molnar, 2022; Ribeiro et al., 2016); 3) the use of personal data in AI raises concerns about privacy and consent (ethical and privacy issues) (Richards, 2021; Véliz, 2021); 4) data quality is critical to AI performance, and faulty data can lead to erroneous results (data quality dependency) (Byabazaire et al., 2020); 5) AI-driven automation can displace human jobs, creating economic and social challenges (unemployment and job displacement) (Acemoglu, et al., 2022; Acemoglu & Restrepo, 2019; Frey & Osborne, 2017) 6) AI can be used for harmful purposes, and AI systems are vulnerable to attack and manipulation (security and misuse) (Brundage et al., 2018) .

to describe the state of the art of the European ethical and legal framework for responding to this phenomenon in an systematic and workable way.

2. DISCRIMINATORY BIAS IN AI

2.1. Scope and impact of discriminatory bias in AI

Within the existing body of literature, comprehensive delineations of the phenomenon known as algorithmic discrimination are infrequent. Instead, comprehension arises predominantly from the ramifications it engenders, especially those entailing inequitable or disparate decision-making among individuals devoid of apparent rationale (O'Neil, 2016; Buolamwini & Gebru, 2018; Eubanks, 2018; Noble, 2018). Consequently, manifestations of discriminatory trends stemming from the deployment of AI tools manifest across diverse domains, including those previously elucidated, along with others necessitating the use of such tools, such as the medical realm (Rajkomar et al., 2018; Obermeyer et al., 2019) and the economic sphere (Mendes & Mattiuzzo, 2022). Such manifestations harbour the potential to result in uneven treatment predicated on factors encompassing race, gender, ethnicity, and more. And, similarly, algorithmic discrimination can also occur when "a computerized model makes a decision or a prediction that has the unintended consequence of denying opportunities or benefits more frequently to members of a protected class than to an unprotected control set" (Brownstein, 2022).

Algorithmic discrimination, in this sense, is the harmful consequences experienced by individuals as a result of outcomes generated by AI tools that operate with specific algorithms. These patterns of discrimination are significant. The need for extensive data collection to support labelling, profiling, recognition or decision making driven by AI algorithms, and the resulting consequences, has sparked a profound debate about the potential impact on individuals. For example, when examining any of these tools, algorithmic profiling often emerges as a source of discrimination (Eubanks, 2018; Mann & Matzner, 2019; Noble, 2018), along with the phenomenon known as the chilling effect (Büchi et al., 2020; FRA, 2019). The chilling effect embodies altered behavioural patterns resulting from fear of surveillance: a form of self-censorship in which individuals strive to avoid negative external perceptions or present an overly positive image. These algorithms work by identifying correlations and making predictions about group-level behaviour, with groups (or profiles) being continually redefined by the algorithm (Zarsky, 2013). Understanding of individuals, whether dynamic or static, is based on associations with others identified by the algorithm, rather than being rooted in actual behaviour (Newell & Marabelli, 2015). As a result, profiling often shapes decisions about individuals through group-derived information (Danna & Gandy, 2002; Malek, 2022), inadvertently leading to the creation of databases that facilitate discrimination (de Vries, 2010). Furthermore, as will be elucidated in the following sections, discriminatory analyses rooted in various types of prejudice can foster self-fulfilling prophecies, misuses, stigmatising marginalised groups and impeding their autonomy and social participation (Cerezo, Roteda-Ruffino & Castro-Toledo, 2021; Leese, 2014; Macnish, 2012).

On the other hand, eminent challenges plaguing AI tools that rely on data training focus on the origin of the data. A significant proportion of algorithmic discrimination arises from non-random patterns within data, derived from pre-existing biased databases. This includes

imbalances in age, gender, ethnicity and other relevant risk factors, as well as outdated or inaccurate data. Similar challenges may arise from analytical shortcomings due to insufficient data or other reasons. However, these challenges may be no more severe than those encountered when human decision making is carried out without computer systems (Bechtel et al., 2017). Nevertheless, there is still a palpable reality: significant ethical challenges remain, stemming from the lack of a well-defined and operational concept of algorithmic fairness. Some characterise this as the need for algorithmic results to be equally accurate, or to produce an equal number of false positives and false negatives for members of different social groups (Hellman, 2020).

Concerns have also arisen about whether the effectiveness and validity of these tools vary according to the gender of the individual being assessed. Specifically, whether the predictive capacity of algorithmic tools remains consistent regardless of gender, or whether effectiveness varies due to the predominantly male-centric composition of validation samples. For example, the Level of Service Inventory-Revised (LSI-R), which is widely used in the United States, has been criticised for its specificity in predicting antisocial male behaviour, with a weaker predictive ability for female behaviour. This has led to calls for the development of more gender-specific instruments (Smith et al., 2009; Olver & Stockdale, 2022) and for gender-sensitive approaches to misconduct risk assessment (Hannah-Moffat, 2009). Contrary findings have also been reported, such as the gender-neutrality of the DRAOR tool as found by Scanlan et al. (2020). Therefore, gender dynamics in risk prediction warrant a comprehensive review that addresses the neutrality of tools in this regard.

2.2. Some key normative-ethical bases for the European response to AI shortcomings

In the previous section, the various reasons for identifying the negative and problematic discriminatory effects of the use of AI tools are manifold and far from being fully addressed. In this regard, the gradual advancement of AI functionalities has led to growing concerns about the ethical, legal and social consequences of their design, development and deployment. These concerns have spurred the creation of numerous ethical and regulatory frameworks in the European context, with the main objective of defining, analysing, minimising and mitigating the potential impacts that AI tools may have in different application contexts.

An examination of the European normative-ethical framework reveals a common consensus, despite possible differences in the interpretation of concepts. This consensus emphasises that AI-driven tools should be developed, deployed and used in accordance with a set of principles, both at an ethical level, including fairness, accountability or transparency, among others; and at a legal level, such as respecting fundamental rights through non-discrimination and the right to privacy. This ethical and legal approach aims to establish a unified European framework for the development of AI. In other words, the incorporation of AI in various sensitive areas has potential implications for fundamental rights and civil liberties if clear limits are not set for its use (FRA, 2018a; FRA, 2018b; FRA, 2018c). Therefore, an AI system that complies with a set of ethical and legal standards underpins the entire rationale of European research and aspirations, and AI designers should respect fundamental European ethical values such as justice, fairness, privacy or transparency for different ethical and

normative reasons. Here only a few will be mentioned. First, because ethical AI can help avoid the emergence and spread of biases that lead to discriminatory or stigmatising practices. Training with data that is biased by race, gender, age or other factors can be key to perpetuating and reinforcing existing prejudices and inequalities. To mitigate this problem, it is precisely necessary to develop AI that is unbiased and takes into account principles such as diversity, universality or plurality (STOA & EPRS, 2022; FRA, 2019b; AI HLEG, 2019). Second, ethical AI can contribute to public benefit. AI has the potential to address many global challenges. However, if used unilaterally or against the shared values of society as a whole, it can have potential consequences for both users and those affected by it (FRA, 2022; AI HLEG, 2019). It is therefore prudent to develop AI that contributes to the well-being of society as a whole, not just some groups. Third, ethical AI can foster trust in technology and innovation. The trust that developers, end-users and citizens can place in AI systems is fundamental to their effective and safe use (FRA, 2020; AI HLEG, 2019; STOA & EPRS, 2020). If key actors involved in the development, implementation and use of AI do not trust it, they may be reluctant to use it, which could limit its operability. It is therefore necessary to develop AI that is transparent, responsible, explainable and accountable. Fourth, ethical AI can be safer, more accurate and more reliable. AI tools can be subject to errors, third-party attacks and manipulation, which can have serious consequences for both users and those affected by them. By developing AI that respects ethical standards, more robust security, privacy, accuracy, tuning and monitoring measures can be implemented, which can reduce the risk of security incidents and improve the reliability of the technology (FRA, 2019b; AI HLEG, 2019).

In addition, issues in relation to algorithmic discrimination have also been the focus of attention in some relevant European guidelines, such as “The Ethics of artificial intelligence: Issues and initiatives” by STOA in 2020, the “Ethics Guidelines for Trustworthy AI” by EC experts in 2019, and its modelling by the “Assessment List for Trustworthy AI” (ALTAI), among others. The existence of potential biases in AI tools is a serious concern and a source of analysis and discussion on their definition, impact and strategies to minimise and mitigate them. Likewise, it does not seem possible to limit the existence of these biases to a specific phase of the overall development of the tools, but rather it is a cross-cutting problem that can be present and affect both the designers and the end users of these tools throughout the process. It is precisely this approach that is reflected in the concept of Ethics by Design, also developed in the European reference framework by the EC in the document, “Ethics by Design and Ethics of Use Approaches for Artificial Intelligence” (2021a). Following the definition given for this approach: "Ethics by Design is an approach that can be used to ensure that ethical requirements are properly addressed during the development of an AI system or technique" (EC, 2021a, p.11) and that: "Ethics by Design aims to prevent ethical issues from arising in the first place by addressing them during the development phase, rather than trying to fix them later in the process" (EC, 2021a, p.12).

Regardless of the methodologies used to carry out this monitoring, which may vary depending on the starting conditions or expected uses, the central point is that the focus should not be exclusively on mitigation measures to address the impacts that may be caused by the misuse of AI tools. Instead, it may be more beneficial to adopt an approach based on prevention of problems that are already recognised as existing and having a potential impact on people. For example, a detailed analysis of potential risks in the form of biases that

developers may face in the early stages of ideation and development of AI tools may ultimately lead to a reduction in potential harmful impacts on people. Or, for example, it could be precautionary to analyse the databases used for algorithmic decision-making at the source, as also stated in the STOA 2022 report, "Auditing the quality of datasets used in algorithmic decision-making systems", which, as has been pointed out, can be a clear source of bias from the design of the data, its collection, processing and maintenance.

In this sense, the establishment of a permanent monitoring task, covering all phases from design and implementation to end-user use, could significantly improve the development of these AI tools. These European guidelines include recommendations and strategies, both at the ethical and legal level, to achieve reliable AI and also to avoid possible biases arising from its development or use. In any case, although these recommendations and practices aimed at analysing the impact and establishing recommendations to minimise and mitigate bias are present in all European documents, ALTAI is currently the guideline that establishes the clearest and most concise way to address them, as it sets out specific questions regarding their possible impact and the specific actions to deal with them, in terms of avoiding unfair bias, accessibility and universal design, and stakeholder participation.

To sum up, in the European context, the wide range of concerns about the ethical, legal and social implications of AI has led to the development of ethical and regulatory frameworks. These frameworks aim to ensure that AI adheres to ethical principles, respects fundamental rights and addresses potential biases throughout its life cycle. However, their correct assimilation and consequently their correct implementation by all interested parties has been complicated by both the growing number of European guidance attempts and their dispersion over time. In response to this complex context of institutional guidelines, their content will be systematised in the following section.

3. MAPPING OF THE MAIN EUROPEAN ETHICAL AND NORMATIVE AI GUIDELINES

3.1. Methodology applied and AI (or related) guidelines assessed

The review was carried out on the basis of 21 European guidelines issued by different public institutions according to the following inclusion criteria:

- Date, establishing a timeline that oscillates between 2016 with the first document considered (GDPR) and 2023 with the las update on the Artificial Intelligence Act (AI Act).
- Topic, namely: AI, bias, algorithmic biases, discrimination, algorithmic discrimination.
- Published by a public European institution, such as: European Union Agency for Fundamental Rights (FRA), Panel on the Future of Science and Technology (STOA), European Parliamentary Research Service (EPRS), European Commission (EC), High-Level Expert Group on Artificial Intelligence (AI HLEG), Council of the European Union (CoEU), Council of Europe (CoE) and European Parliament (EP).

Table 1 provides a chronology of the main ethical and regulatory guidelines that marked the progress of AI in Europe from 2016 to 2023 included in the analyses.

Table 1. Chronology of the main EU ethical and normative AI guidelines analysed.

Date	Document	Institution
2023	Artificial Intelligence Act	EC, EP, CoEU
2022	Bias in Algorithms – Artificial Intelligence and Discrimination	FRA
	Auditing the quality of datasets used in algorithmic decision-making systems	STOA & EPRS.
2021	Ethics by design and Ethics of Use Approaches for Artificial Intelligence	EC
	Algorithmic discrimination in Europe. Challenges and opportunities for gender equality and non-discrimination law	EC
2020	Getting the Future Right. Artificial Intelligence and Fundamental Rights	FRA
	Presidency conclusions -The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change	CoEU
	Assessment List for Trustworthy AI (ALTAI)	EC (AI HLEG)
	Recommendation CM/Rec(2020)1 of the Committee of Minister to member States on the human rights impacts of algorithmic systems	CoE
	The Ethics of artificial intelligence: Issues and initiatives	STOA & EPRS
	Gender Equality Strategy 2020-2025	EC
	White Paper on artificial intelligence -A European approach to excellence and trust	EC
2019	Data quality and artificial intelligence -mitigating bias and error to protect fundamental rights-	FRA
	Unboxing artificial intelligence: 10 steps to protect human rights	CoE
	Ethics Guidelines for Trustworthy AI (HLEG)	EC (AI HLEG)
	Understanding algorithmic decision-making: Opportunities and challenges	STOA & EPRS
2018	Preventing unlawful profiling today and in the future: a guide	FRA
	BigData: Discrimination in data-supported decision making	FRA
	European AI Strategy	EC
2017	Fundamental rights implications of big data	EP
2016	General Data Protection Regulation (GDPR)	EP & CoEU

On the other hand, the following variables were systematically evaluated in each of the included document and answered in a dichotomous way (i.e., yes or no):

- whether or not they provide a definition of the term bias or algorithmic bias,
- whether they establish recommendations or measures to mitigate and minimise algorithmic bias;
- whether or not they are in force. In the case of reports that cannot be directly implemented, the answer “NO" has been chosen to indicate that these are guidelines that can be used for analysis but, strictly speaking, are documents whose content is not mandatory.

3.2. General overview

The review of European guidelines on AI reveals some key findings (Table 2, Figure 1). While the concept of bias is seldom explicitly mentioned, the documents acknowledge its origins and multifaceted impacts—social, legal, and ethical. Only 19% of the documents directly address bias, with 81% omitting it. Furthermore, 86% of the documents propose various measures to mitigate bias, with 14% lacking such measures. Notably, only 10% of the analysed documents and regulations are legally binding, while the remaining 90% are non-binding as most are informative reports, briefings, or studies offering guidance and recommendations, rather than binding directives for member States.

Table 2. Qualitative summary of European ethical and normative AI guidelines

Reference	Date of publication (dd/mm/yy)	Scope of the document	Definition of discriminatory bias	Mitigation / minimisation measures	Legally binding
Artificial Intelligence Act (AI Act)	21/04/2021	It's a proposed European law on artificial intelligence (AI). The law assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk. Second, high-risk applications. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.	NO	YES	YES
Bias in Algorithms – Artificial Intelligence and Discrimination	8/12/2022	The report looks at the use of artificial intelligence in predictive policing and offensive speech detection. It demonstrates how bias in algorithms appears, can amplify over time and affect people's lives, potentially leading to discrimination. It corroborates the need for more comprehensive and thorough assessments of algorithms in terms of bias before such algorithms are used for decision-making that can have an impact on people.	YES	YES	NO
Auditing the quality of datasets used in algorithmic decision-making systems	25/07/2022	This study begins by providing an overview of biases in the context of artificial intelligence, and more specifically to machine-learning applications. The second part is devoted to the analysis of biases from a legal point of view. The analysis shows that shortcomings in this area call for the implementation of additional regulatory tools to adequately address the issue of bias. Finally, this study puts forward several policy options in response to the challenges identified.	NO	YES	NO
Ethics by design and Ethics of Use Approaches for Artificial Intelligence	25/11/2021	Offers guidance for adopting an ethically-focused approach while designing, developing, and deploying and/or using AI based solutions. It explains the ethical principles which AI systems must support and discusses the key characteristics that an AI-based system/ applications must have in order to preserve and promote.	NO	YES	NO
Algorithmic discrimination in Europe. Challenges and opportunities for gender equality and non-discrimination law	10/03/2021	This report investigates how algorithmic discrimination challenges the set of legal guarantees put in place in Europe to combat discrimination and ensure equal treatment. More specifically, it examines whether and how the current gender equality and non-discrimination legislative framework in place in the EU can adequately capture and redress algorithmic discrimination.	YES	YES	NO
Getting the Future Right. Artificial Intelligence and Fundamental Rights	14/12/2020	This report presents concrete examples of how companies and public administrations in the EU are using, or trying to use, AI. It focuses on four core areas – social benefits, predictive policing, health services and targeted advertising.	NO	NO	NO
Presidency conclusions -The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change	21/10/2020	Conclusions on the charter of fundamental rights in the context of artificial intelligence and digital change. These conclusions are designed to anchor the EU's fundamental rights and values in the age of digitalisation, foster the EU's digital sovereignty and actively contribute to the global debate on the use of artificial intelligence with a view to shaping the international framework.	NO	YES	NO
Assessment List for Trustworthy AI (ALTAI)	17/07/2020	Through the Assessment List for Trustworthy AI (ALTAI), AI principles are translated into an accessible and dynamic checklist that guides developers and deployers of AI in implementing such principles in practice. ALTAI will help to ensure that users benefit from AI without being exposed to unnecessary risks by indicating a set of concrete steps for self-assessment.	YES	YES	NO
Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems	8/04/2020	Underlying that member States must ensure that any design, development and ongoing deployment of algorithmic systems occur in compliance with human rights and fundamental freedoms, which are universal, indivisible, interdependent and interrelated, with a view to amplifying positive effects and preventing or minimising possible adverse effects.	NO	YES	NO
The ethics of artificial intelligence: Issues and initiatives	11/03/2020	The study deals with the ethical implications and moral questions that arise from the development and implementation of artificial intelligence (AI) technologies. It also reviews the guidelines and frameworks that countries and regions around the world have created to address these. It presents a comparison between the current main frameworks and the main ethical issues, and highlights gaps around mechanisms of fair benefit sharing; assigning of responsibility; exploitation of workers; energy demands in the context of environmental and climate changes; and more complex and less certain implications of AI, such as those regarding human relationships.	NO	YES	NO
Gender Equality Strategy 2020-2025	5/03/2020	This Gender Equality Strategy frames the European Commission's work on gender equality and sets out the policy objectives and key actions for the 2020-2025 period.	NO	NO	NO
White Paper on artificial intelligence-A European approach to excellence and trust	19/02/2020	The document gives a definition of AI, underlining it's benefits and technological advances in different areas, including medicine, security, farming, as well as identifying it's potential risks: opaque decision making, gender inequality, discrimination, lack of privacy, bias, etc.	NO	YES	NO
Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights	11/06/2019	Algorithms used in machine learning systems and artificial intelligence (AI) can only be as good as the data used for their development. High quality data are essential for high quality algorithms. Yet, the call for high quality data in discussions around AI often remains without any further specifications and guidance as to what this actually means.	NO	YES	NO
Unboxing artificial intelligence: 10 steps to protect human rights	14/05/2019	The document provides a number of steps which national authorities can take to maximise the potential of artificial intelligence systems and prevent or mitigate the negative impact they may have on people's lives and rights. It focuses on 10 key areas of action.	NO	YES	NO
Ethics Guidelines for Trustworthy AI (HLEG)	08/04/2019	The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy.	YES	YES	NO
Understanding algorithmic decision-making: Opportunities and challenges	05/03/2019	The expected benefits of Algorithmic Decision Systems (ADS) may be offset by the variety of risks for individuals (discrimination, unfair practices, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.) and society as a whole (manipulation, threat to democracy, etc.). They present existing options to reduce the risks related to ADS and explain their limitations. They sketch some recommendations to overcome these limitations to be able to benefit from the tremendous possibilities of ADS while limiting the risks related to their use. Beyond providing an up-to-date and systematic review of the situation, the report gives a precise definition of a number of key terms and an analysis of their differences. The main focus of the report is the technical aspects of ADS. However, other legal, ethical and social dimensions are considered to broaden the discussion.	NO	YES	NO
Preventing unlawful profiling today and in the future: a guide	05/12/2018	This guide explains what profiling is, the legal frameworks that regulate it, and why conducting profiling lawfully is both necessary to comply with fundamental rights and crucial for effective policing and border management. The guide also provides practical guidance on how to avoid unlawful profiling in law enforcement agencies and border management operations.	NO	YES	NO
BigData: Discrimination in data-supported decision making	30/05/2018	This focus paper specifically deals with discrimination, a fundamental rights area particularly affected by technological developments.	NO	YES	NO
European AI Strategy	25/04/2018	Aims at making the EU a world-class hub for AI and ensuring that AI is human-centric and trustworthy.	NO	YES	NO
Fundamental rights implications of big data	14/03/2017	The text considers the potential use of big data in both commercial and law enforcement areas, as well as the risks, particularly in terms of unlawful discrimination and bias. It also emphasises the need for greater algorithmic accountability and transparency, calling on the Commission and Member States to ensure, with appropriate guidelines, that data-driven technologies do not jeopardise the exercise of fundamental rights.	NO	YES	NO
General Data protection Regulation (GDPR)	27/04/2016	The general data protection regulation (GDPR) protects individuals when their data is being processed by the private sector and most of the public sector. The processing of data by the relevant authorities for law-enforcement purposes is subject to the data protection law enforcement directive (LED) instead. No mention of biases.	NO	NO	YES

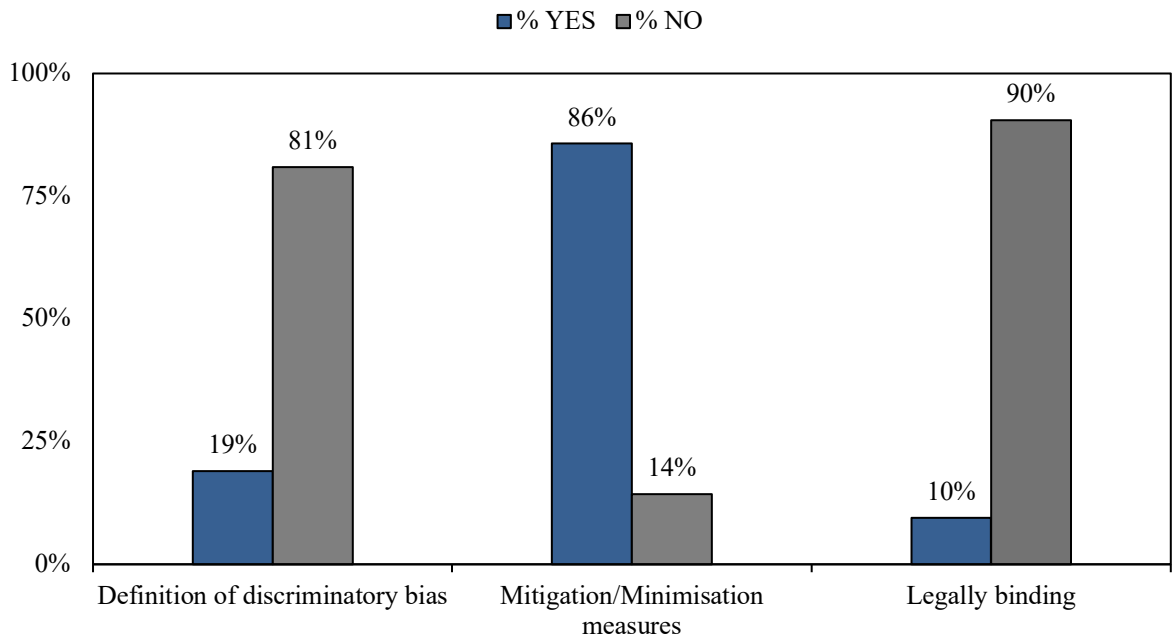

Figure 1. Quantitative summary of the analysis of European documents regarding the ethical and regulatory framework of AI

3.3. Similarities and differences in definitions of algorithmic bias

With regard to the definition of the phenomenon of algorithmic bias, only four guidelines provide an explicit definition (see Annex I). In particular, this section discusses the similarities and differences between the definitions given in 1) "Bias in Algorithms – Artificial Intelligence and Discrimination" (FRA, 2022), 2) "Algorithmic discrimination in Europe" (EC, 2021c), 3) "Assessment List for Trustworthy AI" (ALTAI) (AI HLEG, 2020), and "Ethics Guidelines for Trustworthy AI" (AI HLEG, 2019) ". All definitions exhibit commonalities as they acknowledge that algorithmic bias within AI systems has the capacity to result in unjust or discriminatory outcomes. Whether it takes the form of differential treatment rooted in protected characteristics, systematic errors, or instances of unfairness, there is a shared consensus concerning the potential adverse effects. Additionally, there is unanimous agreement across these definitions that bias in AI can originate from a multitude of sources encompassing data handling, algorithm design, and societal norms. This collective recognition underscores the intricate and multifaceted nature of the issue at hand. Furthermore, each of these definitions acknowledges that algorithmic bias is not confined to a mere technical interpretation but rather embraces a multidimensional concept that necessitates consideration of various facets, ranging from the technical intricacies involved to the ethical implications it carries.

Differences among these definitions become evident when considering their respective emphases. The definition found in "Bias in Algorithms – Artificial Intelligence and Discrimination (FRA, 2022)" places primary focus on the legal and normative dimensions of bias, with a particular emphasis on discrimination and bias-motivated crimes, distinguishing it from the others that encompass a more extensive range of technical and ethical

considerations. "Algorithmic discrimination in Europe" (EC, 2021c) introduces a notable distinction between general systematic errors and those specifically tied to fairness, a subtle nuance absent from the remaining definitions, thereby underscoring the significance of fairness as a distinct facet within the realm of algorithmic bias. On the other hand, the definitions offered by "Assessment List for Trustworthy AI (ALTAI) (AI HLEG, 2020)" and "Ethics Guidelines for Trustworthy AI (AI HLEG, 2019)" accentuate the diversity of AI platforms and systems in which bias may emerge, implying a broader applicability than the initial two definitions. This expansive viewpoint acknowledges that bias can manifest across an array of AI contexts and systems, emphasising its multifaceted presence in the AI landscape.

In summary, while these definitions of algorithmic bias share common ground in recognising its negative consequences and diverse sources, they also exhibit differences in focus, nuance and breadth of application. These differences reflect the multidisciplinary nature of the concept and the need to address it from different angles, including legal and ethical considerations as well as technical aspects.

3.4. Recommendations to mitigate/minimise algorithmic bias

Finally, this review presents an organised compilation of mitigation/minimisation measures extracted from the 86% of the analysed guidelines (see section 3.2). These measures are intended to serve as practical recommendations for effectively addressing discriminatory biases within AI systems (see Annex II). To facilitate clarity and comprehensiveness, the measures identified in this review have been categorised into three excluding categories depending on the stage of development of the AI systems:

- 1) Design: in terms of technical designing issues of the AI system.
- 2) Governance: during the internal management of the development of the AI system.
- 3) Organisational: regarding the implementation and monitoring of the AI system.

The total number of measures considered in this compilation is 148, encompassing a comprehensive set of approaches for mitigating bias in AI. The mean total score across all categories is 7.05. The standard deviation of 4.25 indicates moderate variability among the measures. The measures in the compilation exhibit a wide range, with scores ranging from 0 to 14 in a single document, reflecting the diversity and complexity of bias mitigation strategies in AI contexts. Table 3 summarises the quantitative results of the compilation of bias mitigation/minimisation measures.

Table 3. Description of quantitative results of bias mitigation measures compilation

Type of measure	*n*	%	*M*	*SD*	*Min*	*Max*
Design	53	36	2.52	1.91	0	8
Governance	62	42	2.95	2.46	0	7
Organisational	33	22	1.57	1.72	0	7
Total	148	100	7.05	4.25	0	14

In order to summarise the bias minimisation/mitigation measures presented in Annex II in a practical way, Figure 2 presents the excluding subcategories that have been established from the original recommendations:

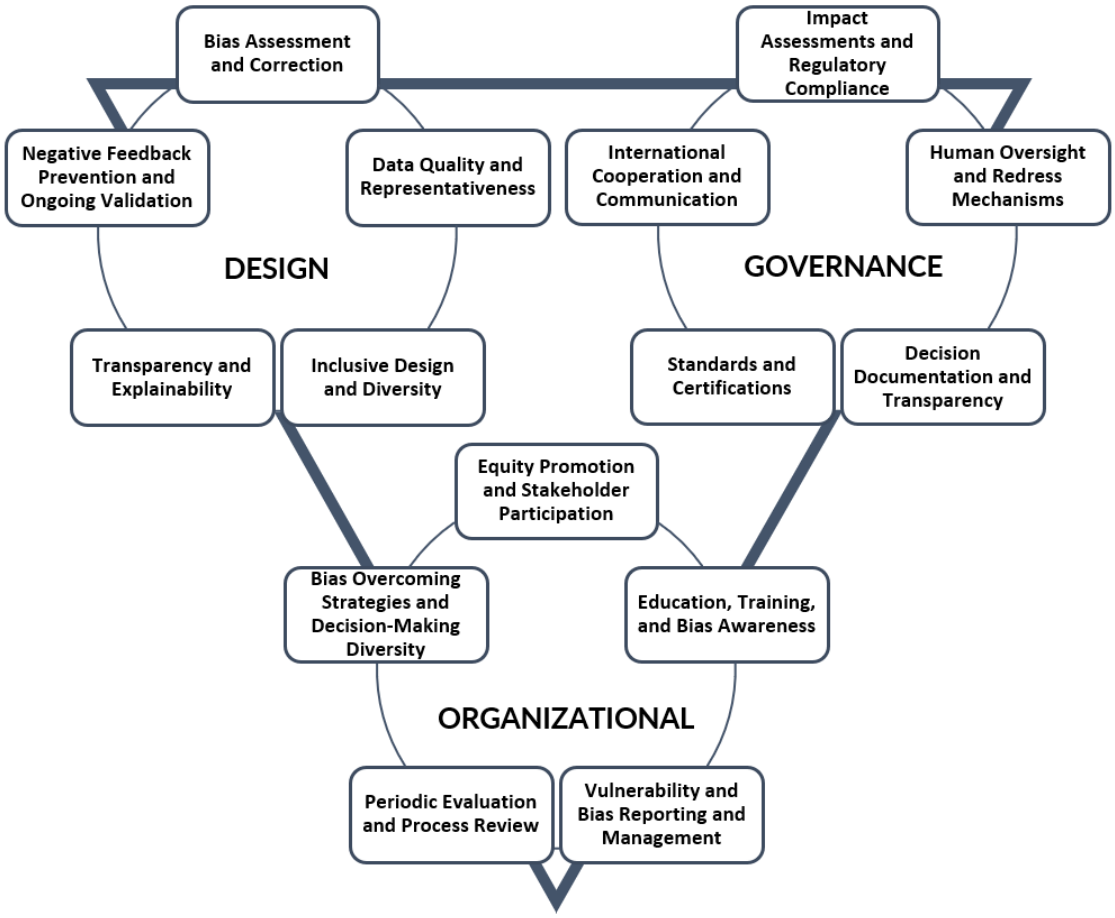

Figure 2. Categories of bias minimisation measures from the European regulatory framework of AI

a) *Design measures*

53 measures have been identified in this category, accounting for 36% of the total. The mean score for Design measures is 2.52. The standard deviation is 1.91, indicating a moderate level of variation. The measures in this category range from a minimum score of 0 to a maximum of 8.

- Bias assessment and correction: directly address the identification and rectification of biases within ai systems. It involves pre-deployment testing for biases, ongoing monitoring, and the implementation of algorithmic adjustments to mitigate identified biases.
- Data quality and representativeness: ensure that the datasets used for training ai systems are accurate, comprehensive, and reflective of the diversity of the target population. This

includes the collection of high-quality data, the assessment of data sources for representativeness, and the elimination of data that may introduce or perpetuate bias.

- Inclusive design and diversity: inclusion of a wide range of linguistic, cultural, and demographic characteristics in the development of ai systems. It promotes the creation of tools and models that can understand and process diverse forms of natural language and cater to a broad user base.
- Transparency and explainability: develop ai systems' ability to provide clear, understandable explanations for its decisions and actions. This includes the development of interpretable models, the documentation of algorithmic processes, and the communication of ai system capabilities and limitations to users.
- Avoidance of feedback loops and ongoing validation: implement mechanisms to prevent ai systems from perpetuating and reinforcing their own biases over time, often referred to as "feedback loops." it also involves the continuous validation of ai systems to ensure they are performing as intended and without discriminatory effects.

b) *Governance measures*

This category comprises 62 measures, making up 42% of the total. The mean score for Governance measures is slightly higher at 2.95. The standard deviation is 2.46, indicating a relatively higher degree of variation compared to Design measures. The range of scores for Governance measures spans from 0 to 7.

- Impact assessments and regulatory compliance: conduct fundamental rights impact assessments and ensure compliance with regulations such as gdpr and existing laws and regulations. They ensure that ai systems are developed and deployed in compliance with legal standards.
- Human oversight and redress mechanisms: ensure human involvement in the oversight of ai systems. It also includes the establishment of mechanisms for individuals to seek redress if they are adversely affected by an ai system.
- Decision-making transparency: record and share key decisions made during the development and deployment of ai systems. They aim to create an audit trail that can be reviewed to ensure ethical and regulatory compliance.
- Standards and certifications: adopt and adhere to industry standards and certifications that guarantee the quality and ethical integrity of datasets and ai mechanisms.
- International cooperation and communication: share best practices, research findings, and policy approaches across international borders. It promotes collaboration among nations, organisations, and stakeholders in the field of ai to establish common standards.

c) *Organizational measures*

33 measures have been identified in the Organizational category, representing 22% of the total. The mean score for Organizational measures is 1.57. The standard deviation is 1.72, suggesting a moderate level of variability. Organizational measures have scores ranging from a minimum of 0 to a maximum of 7.

- Equity promotion and stakeholder participation: promote fairness and equity in ai systems and encourage the involvement of diverse stakeholders throughout the ai lifecycle, from design to deployment and evaluation.
- Education, training, and bias awareness: develop educational programs and training initiatives for ai designers and developers on recognizing and managing biases or potential biases.
- Vulnerability and bias reporting and management: establish protocols for reporting and managing potential vulnerabilities and biases in ai systems. This includes the creation of channels through which internal staff and external parties can report concerns.
- Periodic evaluation and process review: implement regular assessments to ensure data accuracy and representativeness and that ai systems processes continue to function without biases and with accuracy.
- Bias overcoming strategies and decision-making diversity: develop strategies to handle biases and ensure diversity in decision-making teams, reducing the risk of homogenous biased outcomes.

4. CONCLUSIONS

Throughout the previous sections that make up this work, different aspects related to the European normative-ethics response to the algorithmic biases in the context of AI systems have been described. Our study adds significant added value in a number of ways to the state-of-the-art. Overall, the study's first-time combination of European scope, systematic review of authoritative documents, and practical applicability make it a valuable resource for advancing the understanding and management of discriminatory bias in AI systems. In general terms, the problems surrounding this issue are quite clear with regards to the potential negative impacts of biases on certain segments of the population, such as those represented by certain ethnic groups, gender or race. As can be seen from the European guidelines analysed, there is a general trend that systematically disadvantages these groups due to various factors such as faulty data collection, possible biases that the designers of the tools may unconsciously transfer to the algorithm and, finally, the uses that are made of these tools in different contexts. However, there is a heterogeneity of methodologies and definitions when addressing the problem of bias, its impact and its mitigation/minimisation measures. This highlights at least two things: firstly, that there is no clear conceptual framework on the issue at the European level; secondly, that this lack of a clear conceptual framework may affect the concreteness and detail of the potential mitigation/minimisation measures proposed. In other words, if a series of key common definitions are not provided, there may be a lack of clarity and adequacy of the mitigation/minimisation measures to be implemented, preventing them from being more precise depending on the context and the situation. In response to the results of this paper, there is a need for intensification of efforts in some very

interesting lines of research. First, the continuous assessment of a unified European conceptual framework for addressing AI biases, including standard definitions and methodologies (i.e. AI Act). Second, conducting global comparative studies to identify best practices and areas for improvement. Third, advancing technologies to mitigate bias in AI, with a focus on robust and fair algorithms. In addition, studying the specific impact of AI bias in different sectors, such as healthcare and criminal justice, to understand its impact on different populations. Educating and raising awareness of AI biases among developers, policymakers and the public is also crucial. In addition, fair data collection and analysis methods should be explored to minimise inherent biases. Finally, establishing methodologies for regular ethical and social impact assessments of AI systems (beyond ALTAI), with a focus on bias identification and management. These research avenues could significantly improve the understanding and management of discriminatory bias in AI, both in Europe and globally.

REFERENCES

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives*, *33* (2), 3–30. https://doi.org/10.1257/jep.33.2.3

Acemoglu, D., Anderson, G., Beede, D., Buffington, C., Childress, E., Dinlersoz, E., Foster, L., Goldschlag, N., Haltiwanger, K., Kroff, Z., Restrepo, P., & Zolas, N. J. (2022). Automation and the Workforce: A Firm-Level View from the 2019 Annual Business Survey. *NBER Working Paper No. W30659. Social Science Research Network.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4282509

Banu, V. C, Costea, I. M. & Nemtanu, F. C. & Badescu, I. (2017). Intelligent video surveillance system. 2*3RD SIITME*, 208-212. https://doi.org/10.1109/SIITME.2017.8259891

Bechtel, K., Holsinger, A. M., Lowenkamp, C. T., & Warren, M. J. (2017). A meta-analytic review of pretrial research: Risk assessment, bond type, and interventions. *American Journal of Criminal Justice, 42* (2), 443–467. https://doi.org/10.1007/s12103-016-9367-1

Brownstein (2022). Proxy Problems – Solving for Discrimination in Algorithms. *Brownstein Client Alert, feb. 2, 2022*. https://www.bhfs.com/insights/alerts-articles/2022/proxy-problems-solving-for-discrimination-in-algorithms

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Héigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *Apollo - University of Cambridge Repository*. https://doi.org/10.17863/CAM.22520

Büchi, M., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., Velidi S., & Viljoen, S. (2020). The chilling effects of algorithmic profiling: Mapping the issues. *Computer Law & Security Review, 36*. https://doi.org/10.1016/j.clsr.2019.105367

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research, 81*, 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html

Byabazaire, J., O'Hare, G., & Delaney, D. (2020). Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT. *Electronics 2020, 9 (12)*, 2083. https://doi.org/10.3390/electronics9122083

Cerezo-Martínez,P., Roteda-Ruffino, F & Castro-Toledo, F.J. *El reto de los potenciales malos usos de herramientas de IA para uso policial en el I+D europeo* en La transformación algorítmica del Sistema de justicia penal Castro-Toledo, F. J. (Coord.) (2022). https://www.dykinson.com/libros/la-transformacion-algoritmica-del-sistema-de-justicia-penal/9788411254885/

Castro-Toledo, F. J. (Coord.) (2022). *La transformación algorítmica del sistema de justicia penal*. Thomson Reuters Aranzadi. https://www.dykinson.com/libros/la-transformacion-algoritmica-del-sistema-de-justicia-penal/9788411254885/

Council of Europe (2020). *Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems*. Committee of Ministers. https://rm.coe.int/09000016809e1154

Council of Europe (2019). *Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights*. Commissioner for Human Rights. https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights

Council of the European Union (CoEU) (2020). *Presidency Conclusions - The charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change*. https://www.consilium.europa.eu/en/press/press-releases/2020/10/21/artificial-intelligence-presidency-issues-conclusions-on-ensuring-respect-for-fundamental-rights/

Danna, A. & Gandy, O.H. (2002). All That Glitters is Not Gold: Digging Beneath the Surface of Data Mining. *Journal of Business Ethics, 40* (4), 373-386. https://doi.org/10.1023/A:1020845814009

Dastin, J. (2018, October 11). *Insight – Amazon scraps secret AI recruiting tool that showed bias against women*. *Reuters*. https://www.reuters.com/article/idUSKCN1MK0AG/

de Vries, K. (2010). Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology*, *12*(1), 71–85. https://doi.org/10.1007/s10676-009-9215-9

Eubanks, V. (2018). *Automating Inequality: HowHigh-Tech Tools Profile, Law enforcement agencies and Punish the Poor*. London: St. Martin'sPress. https://dl.acm.org/doi/10.5555/3208509

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

2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

European Commission (2021a). *Ethics by design and Ethics of Use Approaches for Artificial Intelligence.* https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

European Commission (2021b). *Laying Down Harmonised Rules on Artificial Intelligence (AI Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Commission (2021c). *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law.* https://op.europa.eu/en/publication-detail/-/publication/082f1dbc-821d-11eb-9ac9-01aa75ed71a1/language-en

European Commission (2020a). *A Union of Equality: Gender Equality Strategy 2020-2025.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0152

European Commission (2020b). *White Paper on Artificial Intelligence.* https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission (2018). *European AI Strategy*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN

European Parliament (2017). *Resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement*. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0076_EN.html

European Union Agency for Fundamental Rights (FRA). (2022). *Bias in algorithms. Artificial intelligence and discrimination*. http://fra.europa.eu/en/publication/2022/bias-algorithm

European Union Agency for Fundamental Rights (FRA). (2020). *Getting the Future Right. Artificial Intelligence and Fundamental Rights*. http://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights

European Union Agency for Fundamental Rights (FRA). (2019a). *Facial recognition technology: fundamental rights considerations in the context of law enforcement*. https://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law

European Union Agency for Fundamental Rights (FRA). (2019b). *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*. https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect

European Union Agency for Fundamental Rights (FRA). (2018a). *Big Data: Discrimination in data-supported decision making*.

https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making

European Union Agency for Fundamental Rights (FRA). (2018b). *Preventing unlawful profiling today and in the future: a guide.* https://fra.europa.eu/en/publication/2018/preventing-unlawful-profiling-today-and-future-guide

European Union Agency for Fundamental Righrs (FRA). (2018c). *Under watchful eyes: biometrics, EU IT systems and fundamental rights*. https://fra.europa.eu/en/publication/2018/under-watchful-eyes-biometrics-eu-it-systems-and-fundamental-rights

Ferguson, A. G. (2017). Policing predictive policing. *Washington University Law Review, 94* (5), 1109-1189. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765525

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Hannah-Moffat, K. (2009). Gridlock or mutability: Reconsidering "gender" and risk assessment. *Criminology and Public Policy*, *8*(1), 209–219. https://doi.org/10.1111/j.1745-9133.2009.00549.x

Heikkilä, M. (2022, March 29). *Dutch scandal serves as a warning for Europe over risks of using algorithms*. Politico. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Hellman, D. (2020). Measuring Algorithmic Fairness. *Virginia Law Review, 106* (4), 811-866. https://virginialawreview.org/articles/measuring-algorithmic-fairness/

Henley, (2021, January 14). *Dutch government faces collapse over child benefits scandal*. The Guardian. https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal

High-Level Expert Group on AI (AI HLEG). (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

High-Level Expert Group on AI (AI HLEG). (2019). *Ethics Guidelines for Trustworthy AI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Leese, M. (2014). The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue, 45*(5), 494–511. https://doi.org/10.1177/0967010614544204

Macnish, K. (2012). Unblinking eyes: the ethics of automating surveillance. *Ethics and Information Technology 14* (2), 151-167. https://doi.org/10.1007/s10676-012-9291-0

Malek, M. A. (2022). Criminal courts' artificial intelligence: the way it reinforces bias and discrimination. *AI Ethics* **2**, 233–245. https://doi.org/10.1007/s43681-022-00137-9

Mann, M., & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society, 6*(2). https://doi.org/10.1177/2053951719895805

Mayson, S. (2019). Bias In, Bias Out. *Yale Law Journal, 128* (35), 2218-2300. https://ssrn.com/abstract=3257004

Mendes, L. S., Mattiuzzo, M. (2022). Algorithms and Discrimination: The Case of Credit Scoring in Brazil. In M. Albers, I. W. Sarlet (Eds.), *Personality and Data Protection Rights on the Internet. Ius Gentium: Comparative Perspectives on Law and Justice, vol 96*. Springer, Cham. https://doi.org/10.1007/978-3-030-90331-2_17

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)*. https://christophm.github.io/interpretable-ml-book/

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification.' *Journal of Strategic Information Systems*, *24*(1), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown. https://dl.acm.org/doi/10.5555/3002861

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366* (6464), 447-453. https://doi.org/10.1126/science.aax2342

Olver, M. & Stockdale, K. C. (2022). Can "Gender Neutral" Risk Assessment Tools be Used with Women and Girls? If so, How?. In S. L. Brown & L. Gelsthorpe (Eds.), *The Wiley Handbook on What Works with Girls and Women in Conflict with the Law: A Critical Review of Theory, Practice and Policy*. John Wiley & Sons. (pp. 102-119). https://doi.org/10.1002/9781119874898.ch8

Panel for the Future of Science and Technology (STOA) & European Parliamentary Research Service (EPRS). (2022). *Auditing the quality of datasets used in algorithmic decision-making systems*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729541

Panel for the Future of Science and Technology (STOA) & European Parliamentary Research Service (EPRS). (2020). *The ethics of artificial intelligence: Issues and initiatives*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452

Panel for the Future of Science and Technology (STOA) & European Parliamentary Research Service (EPRS). (2019). *Understanding algorithmic decision-making: Opportunities and challenges*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624261

Rademacher, T. (2020). Artificial Intelligence and Law Enforcement. In T. Wischmeyer, T. Rademacher (Eds.) *Regulating Artificial Intelligence*. Springer, Cham (pp. 225-254). https://doi.org/10.1007/978-3-030-32361-5_10

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169* (12), 866-872. https://doi.org/10.7326/m18-1990

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, GDPR). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Richards, N. (2021). Why Privacy Matters: An Introduction. *Social Science Research Network. Oxford Press 2021*. https://dx.doi.org/10.2139/ssrn.3973131

Scanlan, J. M., Yesberg, J. A., Fortune, C.-A., & Polaschek, D. L. L. (2020). Predicting women's recidivism using the dynamic risk assessment for offender re-entry: Preliminary evidence of predictive validity with community-sentenced women using a "gender-neutral" risk measure. *Criminal Justice and Behavior, 47*(3), 251–270. https://doi.org/10.1177/0093854819896387

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology and Public Policy*, *8*(1), 183–208. https://doi.org/10.1111/j.1745-9133.2009.00551.x

The White House (n.d.). *Algorithmic Discrimination Protections*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/algorithmic-discrimination-protections-2/#:~:text=Algorithmic%20discrimination%20occurs%20when%20automated,orientation)%2C%20religion%2C%20age%2C

Véliz, C. (2020). *Privacy is Power. Why and How You Should Take Back Control of Your Data*. London: Penguin Random House (Bantam Press). https://www.penguin.co.uk/books/442343/privacy-is-power-by-carissa-veliz/9780552177719

Zarsky, T. Z. (2013). Transparent Predictions. *Illinois Law Review*, 4, 1503–1570. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2324240

ANNEX I. Definitions of algorithmic bias from European AI regulatory framework.

Reference	Definitions
Bias in Algorithms – Artificial Intelligence and Discrimination (2022) (pp.22—24)	The term 'bias' can have a different meaning depending on the context in which it is used and the discipline it comes from, for example law or computer science. It is therefore important to clarify its meaning in the context of this report. Bias can refer to any of the following. — Differential treatment based on protected characteristics, such as discrimination and bias-motivated crimes. This refers to an inclination for or against a person or group based on protected characteristics, such as ethnic origin, gender, religion, colour or sexual orientation. Discrimination defines a situation in which an individual is disadvantaged in some way on the basis of 'one or multiple protected grounds. Crimes committed with a bias motivation are a particularly severe example of a result of biases against people based on their (assumed) characteristics. Such definitions are often used in legal contexts and the social sciences. — Differentiation. Bias understood in this sense is necessary for the proper functioning of a statistical or machine learning algorithm. For example, a machine learning model that has to differentiate between oranges and pears has to have bias towards labelling round, orange objects as oranges. Such use of bias is mainly found in computer science and machine learning. — Statistical bias. This refers to the systematic difference between an estimated parameter and its true value. Statistical bias exists when data are not adequately measuring what they are intended to measure. For example, gross domestic product per capita is not necessarily a good measure of the standard of living in a country, as it does not account for inequality of income distribution. In addition, data and the resulting statistical estimates may not be representative of the target population. For example, if a sample of the general population contains more men than women, it is said to be biased towards men. Bias is mainly understood in this way in statistics. — Offset from origin. In the context of deep learning, bias is also the name for an estimated parameter. The fixed number indicating the average baseline estimate in the linear weight functions of neural networks is called bias; it is often referred to as a 'constant term' or 'intercept' in classical regression analysis. It is a purely technical term, and as such it is not relevant to the present discussions, although it is used in neural networks. Bias is analysed in the context of discrimination (as a legal and normative concept) in this report. Discrimination is mainly linked to prejudices picked up or enshrined in data but may also be the result of statistical bias.
Algorithmic discrimination in Europe. Challenges and opportunities for gender equality and non-discrimination law (2021) (pp..47-48)	Specifically, 'algorithmic bias' refers to 'a systematic error' of any kind in the outcome of algorithmic operations. Bias therefore has a much wider meaning than discrimination as it is not only concerned with unfair errors but with all kinds of 'systematic' errors, which can include those of a statistical, cognitive, societal, structural or institutional nature. When invoked in the context of 'fairness', however, 'algorithmic bias' refers to a particular type of error that 'places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage'. This definition shares commonalities with the legal definition of discrimination understood as the differential unfavourable treatment of an individual or group or the disproportionately disadvantageous impact of a given measure or policy on a specific group. However, the term 'algorithmic bias' is more encompassing than the legal term 'algorithmic discrimination' as it refers to any kind of disadvantage that could be viewed as ethically or morally wrong. For example, an algorithm that disadvantages low-income groups and privileges people with high incomes could be seen as entailing a form of algorithmic bias from an ethical point of view. From a legal point of view, however, algorithmic discrimination only pertains to the unjustified unfavourable treatment of, or disadvantage experienced by, specific categories of population protected by the law either explicitly (e.g. protected grounds) or implicitly (e.g. general or open-textured non-discrimination clauses). For example, in the context of EU gender equality and non-discrimination law, algorithmic discrimination refers to discrimination based on one of the six grounds explicitly listed in and protected under Article 19 TFEU, that is sex, race or ethnic origin, disability, sexual orientation, religion or belief and age. This is why the term 'algorithmic discrimination' will be used throughout this report to refer to the types of algorithmic bias that are problematic from the point of view of EU gender equality and non-discrimination law.
Assessment List for Trustworthy AI (ALTAI) (2020) (p.23)	AI (or algorithmic) bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others. Bias can emerge due to many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm. Bias can enter into algorithmic systems as a result of pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the software's initial design. AI bias is found across platforms, including but not limited to search engine results and social media platforms, and can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity.
Ethics Guidelines for Trustworthy AI (HLEG) (2019) (p.36)	Bias is an inclination of prejudice towards or against a person, object, or position. Bias can arise in many ways in AI systems. For example, in data-drive AI systems, such as those produced through machine learning, bias in data collection and training can result in an AI system demonstrating bias. In logic-based AI, such as rule-based systems, bias can arise due to how a knowledge engineer might view the rules that apply in a particular setting. Bias can also arise due to online learning and adaptation through interaction. It can also arise through personalisation whereby users are presented with recommendations or information feeds that are tailored to the user's tastes. It does not necessarily relate to human bias or human-driven data collection. It can arise, for example, through the limited contexts in which a system in used, in which case there is no opportunity to generalise it to other contexts. Bias can be good or bad, intentional or unintentional. In certain cases, bias can result in discriminatory and/or unfair outcomes, indicated in this document as unfair bias.

ANNEX II. Bias mitigation measures raised in European AI related documents.

Reference	Year	Type of measure	Specific minimisation/mitigation measures
ARTIFICIAL INTELLIGENCE ACT (AI Act) (pp. 3, 11-16, 26-30, 48, 52)	2021 / 2023	Design	1. Certain high-risk AI systems (such as those whose purpose is to assist judicial authorities in investigating and interpreting facts and law and in applying the law to specific facts) are subject to specific requirements on logging capabilities and human oversight in order to avoid the risk of possible biases, errors and opacities and other technical inaccuracies that lead to biased results. 2. Providers should be able to process also special categories of personal data in order to ensure the bias monitoring, detection and correction in relation to high-risk AI systems (training, validation and testing of data) always providing appropriate safeguards for the fundamental rights and freedoms of natural persons as required by the relevant Directives and Regulations, including technical limitations on re-use and the use of the latest security and privacy protection measures, such as pseudonymisation or encryption, where anonymisation would significantly affect the intended purpose. 3. Develop technical robustness against malicious actions that could lead to safety impacts or negatively affect fundamental rights. 4. High-risk AI systems that continue to learn after market introduction or commissioning shall be developed in such a way that potential biases in output information due to the use of output as input data in future operations ("feedback loop") are adequately addressed by appropriate mitigation measures.
		Governance	1. Follow ex-ante conformity assessment procedures, rules on data and data governance, documentation and recording keeping, transparency and provision of information to users, and human oversight. 2. In case infringements of fundamental rights still happen, effective redress for affected persons shall be made possible by ensuring transparency and traceability of the AI systems coupled with strong ex post controls. 3. Create codes of conduct.
		Organizational	1. Avoid automation bias, which can end up leading to decisions that are harmful and discriminatory to human beings ("human oversight").
BIAS IN ALGORITHMS: ARTIFICIAL INTELLIGENCE AND DISCRIMINATION (pp. 7-15, 77-78)	2022	Design	1. Test for bias before and regularly after deployment. 2. Provide guidance on when and how to collect and safeguard data on sensitive attributes and how to assess training data quality (in order to avoid "feedback loops"). 3. Promote language diversity in tools available for natural language processing.
		Governance	1. Decide when an algorithm cannot be used and should be abandoned. 2. Assess ethnic and gender biases, highlighting potential under- and over-flagging of content. 3. Share the information necessary to assess bias with relevant oversight bodies (equality bodies and data protection authorities, which should employ specialised staff and cooperate with data protection authorities and other relevant oversight bodies). 4. Increase knowledge, awareness and resources for bias testing of algorithms (increase access to resources needed for evidence-based oversight of algorithms, share data and data infrastructures…).
		Organizational	1. Assess outputs, specially on particular groups or areas with little research.
AUDITING THE QUALITY OF DATASETS USED IN ALGORITHMIC DECISION-MAKING SYSTEMS (pp. II-III, 16-40)	2022	Design	1. Adopt a preventing approach (using techniques that correct biases in AI systems from the first stages of the AI tool development process, via pre-processing, in-processing, and post-processing). 2. Differentiate between patterns in the data that represent factual knowledge that we want the AI-based system to learn (e.g., obesity increases colorectal cancer risk) and stereotypes that we want to avoid (e.g., fat people do not have exercise habits). 3. Create or use high quality domain-specific training datasets (ensure that training, validation and testing data sets are sufficiently relevant and representative), and continuously assess the quality and integrity of the data.
		Governance	1. Include the 'human in the loop' during the development process and build diverse, interdisciplinary development teams with ethical reflection and inclusive participation. 2. Consider the GDPR concept of "fairness", and apply DPIAS and IA impact assessments. 3. Adopt standards and certificates applicable to datasets and AI mechanisms, both in terms of the information to be included in a dataset and the types of procedures that will ensure the absence of bias in an IA system. 4. Monitor high-risk AI tools and delimitation of uses according to the assigned risk (through adequate tools, such as dynamic monitoring and providing citizens and NGOs with tools to complain or sue).
		Organizational	1. Strengthen AI-system-subject transparency rights to find the source of biased results.
ETHICS BY DESIGN AND ETHICS OF USE APPROACHES FOR ARTIFICIAL INTELLIGENCE (pp. 3-21)	2021	Design	1. Specify the steps which will be taken to ensure data about people is representative of the target population and reflects their diversity or is sufficiently neutral (document how bias in input data and in the algorithmic design will be identified and avoided). Establish a formal process to guarantee the selection of data for the system will be fair, accurate and unbiased (initial assessment, auditable mechanisms…). It should be assumed that any data gathered is biased, skewed or incomplete until proven otherwise.
		Governance	1. Incorporate ethical principles into the development process. 2. "Data minimisation and data protection should never be leveraged to hide bias or avoid accountability, and these should be addressed without harming privacy rights". 3. Fair impacts: ensure that the AI system does not affect the interests of relevant groups in a negative way, and document methods to identify and mitigate negative social impacts in the medium and longer term. 4. Transparency: address all the relevant ethical issues, such as the removal of bias from a dataset, and keep records of all relevant decisions to allow tracing how ethical requirements have been met. 5. Universal accessibility: design AI systems to be usable by different types of end-users with different abilities.
		Organizational	1. Guarantee that both internal staff and third parties can report potential vulnerabilities, risks or biases, and are aware of the limits of the system.
ALGORITHM DISCRIMINATION IN EUROPE. CHALLENGES AND OPPORTUNITIES FOR GENDER EQUALITY AND NON-DISCRIMINATION LAW (pp. 11, 140-151)	2021	Design	1. Include preventive strategies in the design, training and development phases of the creation of algorithms (equality impact assessments and equality by design strategies offering guidance on the equality law framework to computer and data scientists). 2. Implement technological debiasing strategies to minimise algorithmic discrimination both at the level of data selection, labelling and use, and at the level of algorithmic models themselves. 3. Intervene ex post through the use of screening and auditing algorithms that can detect discrimination. 4. Use open and clean data for training and control purposes.
		Governance	1. Create dedicated monitoring and supervising institutions, both public (EU equality body) and private that promote the use of non-discriminatory algorithms. 2. Create soft-law instruments such as ethical codes, self-regulation practices such as voluntary codes of conduct, recommendations and guidelines, cooperation between data protection agencies and equality bodies and the setting up of public-private alliances. 3. Adopt the draft Horizontal Directive under negotiation at the council since 2008 and an expansive interpretation of the personal scope of EU equality law. 4. Continuously monitor and test high-risk algorithms and their output, set up auditing, labelling and certification mechanisms, and encourage watchdogs and whistleblowers to signal suspicions of algorithmic discrimination. 5. Promote a better representation of all minority groups in the professional communities designing and training algorithms to favour a diversity of perspectives (gender equality, among others). 6. Include active human involvement: human-centred AI or human-in-the-loop systems designed to avoid rubber-stamping, complemented by supervision and consultation mechanisms (chain of control and consultation with users), with a clear allocation of liability and legal responsibility. 7. Facilitate legal redress by increasing transparency (e.g. open data requirements for monitoring purposes, such as access to source codes), explainability and accountability, and by combining different legal tools to foster clear attribution of legal responsibilities, clear remedies, fair rules of evidence, flexible and responsive interpretation and application of non-discrimination concepts.
		Organizational	1. Raise awareness, train and educate about the risks of algorithmic discrimination linked to the use of AI and ways to tackle it among IT specialists but also all relevant professional communities (regulators, judges, economic players and the society at large).
GETTING THE FUTURE RIGHT. ARTIFICIAL INTELLIGENCE AND FUNDAMENTAL RIGHTS	2020	Design Governance Organizational	N/A
	2020	Design	1. Data used to train AI systems have to be accurate and adequate for their purpose, and potential biases have to be addressed.

Reference	Year	Type of measure	Specific minimisation/mitigation measures
PRESIDENCY CONCLUSIONES – THE CHARTER OF FUNDAMENTAL RIGHTS IN THE CONTEXT OF ARTIFICIAL INTELLIGENCE AND DIGITAL CHANGE (pp. 5-14)		Governance	1. Address opacity, complexity, bias, a certain degree of unpredictability and partially autonomous behaviour to ensure the compatibility of automated systems with fundamental rights and to facilitate the enforcement of legal rules. 2. Adopt a human-centric and fundamental rights based approach. 3. Pay special attention to marginalised individuals and groups and those in vulnerable situations. 4. Make public participation easier and more effective. 5. Data protection rules and other legal and ethical norms need to be ensured and appropriate safeguards have to be in place, specially in sensitive matters (mass surveillance, facial recognition systems, hate speech in online platforms…). 6. Make AI systems transparent and explicable.
		Organizational	1. Ensure that decisions based on algorithmic systems are less prone to biased results than human-made decisions, and allow better-targeted individual assistance and treatments, benefitting the whole social community and promoting the social protection and healthcare of vulnerable groups.
ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI) (pp. 5, 16-18, 22)	2020	Design	1. Avoid creating or reinforcing historic unfair bias from the data. 2. Consider diversity and representativeness of data (test for specific target groups or problematic uses). 3. Guarantee mechanisms to ensure fairness in your AI system, and a quantitative analysis or metrics to measure and test the applied definition of fairness. 4. Create a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.
		Governance	1. Perform a prior fundamental rights impact assessment to check whether it potentially negatively discriminates against people (testing and monitoring during development, deployment and use phases, and rectifying measures). 2. Enable inclusion and diversity throughout the entire AI system's life cycle. 3. Make sure AI systems are user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. 4. Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system. 5. Set educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system. 6. Identify the subjects that could potentially be (in)directly affected by the AI system, and consult with the impacted communities or groups. 7. Consult stakeholders (solicit regular feedback even after deployment and long term participation).
		Organizational	1. Establish clear steps and ways of communicating on how and to whom bias issues can be raised. Establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system.
RECOMMENDATION CM/REC(2020)1 OF THE COMMITTEE OF MINISTERS TO MEMBER STATES ON THE HUMAN RIGHTS IMPACTS OF ALGORITHMIC SYSTEMS (pp. 7-8, 12-13)	2020	Design	1. Assess quality of datasets in algorithmic systems, considering human rights and non-discrimination rules that may be affected as a result of the quality of the data that are being put into and extracted from an algorithmic system. Attention should be given to the provenance, shortcomings, and the possibility of inappropriate or decontextualized use of the dataset. Be aware of risks related to the quality, nature, and origin of data used for training their systems, ensuring that errors, bias, and potential discrimination in datasets and models are addressed within the specific context. 2. Ensure that the functioning of the algorithmic systems is tested and evaluated with due regard to the fact that outputs vary according to the specific context in which they are deployed and the size and nature of the dataset that was used to train the system, including with regard to bias and discriminatory outputs. 3. Ensure that testing on personal data of individuals is performed with diverse and sufficiently representative sample populations, ensuring that relevant demographic groups are neither over- nor under-represented, and not draw on or discriminate against any particular demographic group.
		Governance	1. Identify and/or develop appropriate institutional and regulatory frameworks and standards that set benchmarks and safeguards to ensure the compatibility of the design, development and ongoing deployment of algorithmic systems with human rights. 2. Invest in relevant expertise to be available in adequately resourced regulatory and supervisory authorities. 3. Regular testing and continuous evaluation, reporting and auditing against state-of-the-art standards related to completeness, relevance, privacy, data protection, other human rights, unjustified discriminatory impacts and security breaches before, during and after production and deployment, to detect technical errors, legal, social, and ethical impacts. 4. Ensure that the staff involved has sufficiently diverse backgrounds to avoid deliberate or unintentional bias. 5. Follow a standard framework for human rights due diligence to avoid fostering or entrenching discrimination throughout all life-cycles of their systems. Seek to ensure that the design, development, and deployment of their systems do not have direct or indirect discriminatory effects on individuals or groups that are affected by these systems, including on those who have special needs or disabilities or who may face structural inequalities in their access to human rights.
		Organizational	1. Foster democratic participation and general public awareness of the capacity, power and consequential impacts of algorithmic systems. 2. Ensure that the development of algorithmic systems is discontinued if testing or deployment involves the externalisation of risks or costs to specific individuals, groups, populations and their environments. 3. Stop the development of algorithmic systems if human rights impact assessments or testing phases identify significant risks or negative externalities that cannot be mitigated.
THE ETHICS OF ARTIFICIAL INTELLIGENCE: ISUES AND INITIATIVES (pp. 2, 16, 30-36, 47)	2020	Design	1. Develop a fairness definition, define what a fair outcome looks like, and include that in the development process. 2. Assume that biases exist within data and thus within systems built from these data, and strive not to replicate them. 3. Search for training data representative of the task and the different groups.
		Governance	1. Allow the communication about the possible existence of biases. 2. Ensure fairness and transparency through being able to know why an automated program made a particular decision: explainable systems, intentional understanding (through validation, investigation and evaluation of the program during development), and algorithm auditors. 3. Accountability: respect the regulation; and Control: "human in the loop", and "the big red button". 4. Minimise the "black box" nature of machine learning, through codes of conduct and initiatives to spot biases earlier.
		Organizational	N/A
GENDER EQUALITY STRATEGY 2020-2025	2020	Design	N/A
		Governance	
		Organizational	
WHITE PAPER ON ARTIFICIAL INTELLIGENCE (pp. 11-15, 18-24)	2020	Design	1. Avoid faulty and biased training data at the design stage, and create mechanisms to ensure that quality of data is maintained throughout the use of AI. 2. Follow specific requirements and control for certain particular AI applications (remote biometric identification).
		Governance	1. Record the process of data selection, keeping of the data, and documentation on programming, training methodologies and techniques avoiding biases. 2. Human oversight: monitoring, intervention and validation of the outcomes, so that it does not lead to outcomes entailing prohibited discrimination. 3. Enable prior conformity assessments and enhance compliance with legal requirements (and its enforcement). 4. Encourage international cooperation. 5. Inform about the capabilities and limitations of the AI system, and against the "black box effect", both for citizens and researchers.
		Organizational	N/A
DATA QUALITY AND ARTIFICIAL INTELLIGENCE - MITIGATING BIAS AND ERROR TO PROTECT FUNDAMENTAL RIGHTS (p.3, pp.8-9, pp.11-13)	2019	Design	1. Establish a constant assessment to ensure the quality of the data through the following measures: the study of possible errors in the data such as lack of precision, representativeness of the samples of the data collected. 2. Use of the concepts of reliability and validity in the collection and processing of the data to be used. 3. Elaboration of detailed descriptions of the data sets to be used in order to be able to know their contents and to guarantee their quality. 4. Ask questions such as: -What information is included in the data? -Is the information included in the data appropriate for the purpose of the algorithm? -Who is covered in the data? -Who is under-represented in the data? -What is the time frame and geographical coverage of the data collection used for building the application?
		Governance	N/A

Reference	Year	Type of measure	Specific minimisation/mitigation measures
		Organizational	N/A
UNBOXING ARTIFICIAL INTELLIGENCE: 10 STEPS TO PROTECT HUMAN RIGHTS (pp. 12-15)	2019	Design	1. Process the data in a proportionate manner in relation to the legitimate purpose pursued by such processing, and shall reflect at all stages of the processing a fair balance between the interests pursued by the development and deployment of the IA system and the rights and freedoms at stake.
		Governance	N/A
		Organizational	1. Introduce a legislative framework providing adequate safeguards where AI systems are based on the processing of genetic data; personal data relating to criminal offences, criminal proceedings and convictions, and related security measures; biometric data; personal data relating to 'racial' or ethnic origin, political opinions, trade union membership, religious or other beliefs, health or sex life. These safeguards should also provide protection against discriminatory or biased processing of these data. 2. Promotion of AI literacy.
ETHICS GUIDELINES FOR TRUSTWORTHY AI (p.12, pp.17-18, p.27-30, p. 29)	2019	Design	1. Ensure the quality and integrity of the datasets that are collected, processed and subsequently used in AI tools. 2. Avoid unfair biases caused by the use of incorrect, outdated or inaccurate data. 3. Establish monitoring and follow-up measures at different stages of the life cycle of AI tools.
		Governance	1. Guarantee the principle of equity through a fair and equal distribution of benefits and costs, ensuring that individuals and groups are not unfairly biased. 2. Ensure inclusion and diversity throughout the lifecycle of AI systems, encouraging participation and ensuring equal access through inclusive design processes. Seek regular feedback even after the deployment of AI systems and establish mechanisms for long-term stakeholder involvement. 3. Ensure accessibility and universal design, so that systems are user-centred, user-friendly and socially responsive. 4. Introduce mechanisms to enable others to report potential problems related to the existence of bias.
		Organizational	1. Guarantee the principle of equity through a fair and equal distribution of benefits and costs, ensuring that individuals and groups are not unfairly biased. 2. Encourage stakeholder participation in developing and auditing AI systems. 3. Communicate potential or perceived risks such as those related to the possible existence of bias. 4. Introduce processes for workers or external parties to report potential vulnerabilities or biases in the IA system or its application.
UNDERSTANDING ALGORITHM DECISION-MAKING: OPPORTUNITIES AND CHALLENGES (pp. I-VII, p.25, 41, pp. 66-67, p.76)	2019	Design	1. Promotion of the principle of fairness. 2. Utilization of certifications and labels in order to enhance the trust in algorithmic decisions systems. 3. Ensure appropriate creation of datasets. 4. Be aware about possible technical constrains. 5. Avoidance/Mitigation of opacity in AI tools. 6. Re-train data constantly. 7. Pre-processing possible existing bias. 8. Give the possibility to test systems across numerous domains and via numerous methodologies.
		Governance	1. Ensure adequate measures in order to avoid non-discrimination. 2. Give the possibility to test systems across numerous domains and via numerous methodologies.
		Organizational	1. Promotion of the principle of fairness. 2. Utilization of certifications and labels in order to enhance the trust in algorithmic decisions systems. 3. Avoidance/Mitigation of opacity in AI tools.
PREVENTING UNLAWFUL PROFILING TODAY AND IN THE FUTURE: A GUIDE (pp.11-12, p. 22,48,60,72, pp.80-81)	2018	Design	1. Use reliable data based on accuracy, quality or representativeness. 2. Algorithmic profiling that is legitimate, necessary and proportionate. 3. Knowledge of fundamental rights and their application in their given context.
		Governance	1. Be aware of fundamental rights and their application in their given context. 2. Conduct assessments to find out whether there are norms and practices that perpetuate explicit or implicit prejudices and negative stereotypes. 3. Ensure that performance indicators are linked to the prevention of prejudice and stereotypes. 4. Introduce specific courses and/or training sessions focusing on addressing personal and institutional bias and stereotypes.
		Organizational	1. Inform individuals by providing them with information about the data to be collected, stored and processed. 2. Be aware of fundamental rights and their application in their given context. 3. Reflect on whether their decision is justified by objective information in order to avoid unlawful or biased profiling. 4. Provide timely and detailed information to officers, for example in 'pre-shift briefings' at the beginning of each shift in order to guide officers on how to conduct their duties. 5. Conduct assessments to find out whether there are norms and practices that perpetuate explicit or implicit biases and negative stereotypes. 6. Introduce specific training courses and/or sessions focused on addressing personal and institutional biases and stereotypes. 7. Ensure that performance indicators are linked to the prevention of prejudice and stereotypes.
BIGDATA: DISCRIMINATION IN DATA-SUPPORTED DECISION MAKING (p.5, 8,11)	2018	Design	1. Highlight the importance of data quality and its potential to affect unfair biases 2. Exclude information about protected groups such as gender or ethnicity from the dataset. 3. Check whether protected characteristics of individuals can be inferred from other information in the dataset, so-called proxies. 4. Ensure that the way the algorithm was constructed and works can be explained in a meaningful way
		Governance	1. Conduct fundamental rights impact assessments: identify possible biases and abuses in the application and results of algorithms.
		Organizational	1. Conduct fundamental rights impact assessments: identify possible biases and abuses in the application and results of algorithms. 2. Ensure that the way the algorithm was constructed and works can be meaningfully explained.
EUROPEAN AI STRATEGY (p.14,16)	2018	Design	1. Develop AI systems in a way that allows humans to understand (the basis for) their actions.
		Governance	1. Supporting research into the development of explainable AI.
		Organizational	1. Develop AI systems in a way that allows humans to understand (the basis for) their actions. 2. Supporting research into the development of explainable AI.
FUNDAMENTAL RIGHTS IMPLICATIONS OF BIG DATA (Articles, 20,21,22; statement M)	2017	Design	1. Establish procedures that can ensure data quality and avoid biased algorithms, spurious correlations, errors or underestimation of legal, social and ethical implications.
		Governance	N/A
		Organizational	1. Establish periodic assessments of the representativeness of data sets, consider whether they are affected by biased elements, and develop strategies to overcome such biases. 2. Review the accuracy and significance of data analysis predictions on the basis of impartiality and ethical concerns. 3. Assess the need not only for algorithmic transparency, but also for transparency about possible biases in the training data used to make inferences based on big data.
GENERAL DATA PROTECTION REGULATION (GDPR)	2016	Design	N/A
		Governance	
		Organizational	

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

All the authors have contributed equally to this work.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101021797