BIAS AND DISCRIMINATION IN ARTIFICIAL INTELLIGENCE SYSTEMS, CAUSES AND LEGAL RESPONSES IN THE EUROPEAN UNION

Sesgos y discriminación en los sistemas de inteligencia artificial: causas y respuestas legales en la Unión Europea

ESTER AVENTIN CASANOVA
Universidad Nacional de Educación a Distancia
ester.aventin15@outlook.es

Cómo citar/Citation
Aventin Casanova, E. (2025).
Bias and discrimination in artificial intelligence systems, causes and legal responses in the European Union.
IgualdadES, 13, 181-208
doi: https://doi.org/10.18042/cepc/IgdES.13.07

(Recepción: 29/08/2025; aceptación tras revisión: 06/10/2025; publicación: 18/12/2025)

Abstract

Artificial Intelligence (AI) increasingly shapes daily life in areas like healthcare, education, and public services. The European Union's Artificial Intelligence Act (Regulation (EU) 2021/0106, 2024) recognizes AI as a powerful tool but highlights risk that require regulation. A major concern is discriminatory bias in AI systems, often caused by poor data quality, misuse, underrepresentation of relevant information in algorithm design, or excessive irrelevant data leading to biased or discriminatory decisions. This paper explores the root causes of algorithmic bias and reviews the EU's legal responses, focusing on the AI Act's regulatory framework. Ensuring fairness and non-discrimination in AI is crucial to protect fundamental rights in an increasingly automated world.

Keywords

Artificial intelligence; algorithmic bias; discrimination; data quality; European Union; Artificial Intelligence Act; fundamental rights; non-discrimination; automated decision-making.

Resumen

La inteligencia artificial (IA) influye cada vez más en la vida diaria en sectores como la sanidad, la educación y los servicios públicos. El Reglamento de Inteligencia Artificial de la Unión Europea (Reglamento (UE) 2021/0106, 2024) reconoce que la IA es una herramienta poderosa, pero con riesgos que deben ser regulados. Uno de los principales es el sesgo discriminatorio, causado frecuentemente por la mala calidad de los datos, el uso indebido, la infrarrepresentación de información relevante en el diseño algorítmico o el exceso de datos irrelevantes que generan decisiones sesgadas o discriminatorias. Este trabajo analiza las causas del sesgo y revisa las respuestas legales de la UE, con énfasis en el marco regulador del Reglamento. Garantizar la equidad y la no discriminación en la IA es clave para proteger derechos fundamentales en un mundo cada vez más automatizado.

Palabras clave

Inteligencia artificial; sesgo algorítmico; discriminación; calidad de los datos; Unión Europea; Reglamento de Inteligencia Artificial; derechos humanos; no discriminación; toma de decisiones automatizada.

SUMMARY

I. ARTIFICIAL INTELLIGENCE AND DISCRIMINATION: ORIGINS AND KEY FACTORS

1. WHAT IS ARTIFICIAL INTELLIGENCE?

In the current legal and technological landscape, it is essential to turn to authoritative sources in order to understand the scope, nature and definition of artificial intelligence. One of the main reference sources in this context comes from legal reports, such as those prepared by the European Union Agency for Fundamental Rights (2020), they define AI as systems that exhibit intelligent behaviour, capable of analysing their environment and taking action —with a certain degree of autonomy— in order to achieve specific objectives. This definition highlights the autonomous nature of AI systems, highlighting their ability to make decisions based on information they receive from the environment, making them key tools for efficiently addressing complex problems. The European Commission's High-Level Expert Group on Artificial Intelligence (2019: 4) underscores the transformative potential of artificial intelligence, considering it not only an end in itself but also acts as a means to promote human prosperity. In this sense, AI is presented as a driver of progress, innovation, and improvement of both individual and social well-being, in addition to contributing to the common good through its

implementation in diverse areas such as health, education, security, and the economy.

In turn, the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021: 4) offers a complementary perspective, understanding AI as "systems capable of processing data and information in a manner that emulates intelligent behaviour", which involves key aspects such as reasoning, learning, perception, prediction, planning, and control. This definition broadens the horizon of AI, emphasizing its practical applications and its ability to perform tasks that, at one time, could only be performed by humans.

In more technical terms, AI can be defined as the ability of a computer system or machine to imitate human cognitive functions, such as reasoning, problem-solving, language comprehension, and learning from data (Russel and Norvig, 2016: 19-23). Stuart Russell and Peter Norvig, two of the most renowned experts in the field, approach AI from a perspective centred on "intelligent agents". In their work *Artificial Intelligence: A Modern Approach*, they describe AI as the study of agents that receive perceptions from the environment and execute actions, emphasizing the importance of the rationality of these systems when making decisions based on the data they receive (Russel and Norvig, 2016: 54-57). This definition highlights the ability of AI systems to act autonomously, in order to achieve a specific goal, without direct human intervention.

This section does not seek to provide a single definition of artificial intelligence, as such a definition must necessarily evolve over time as technological advances continue to shape the field. AI, being a dynamic area of research, requires a flexible approach that allows for the incorporation of new perspectives and characteristics that emerge as new developments emerge (UNESCO, 2021:4).

2. HOW DOES IT WORK?

Understanding artificial intelligence requires not only a conceptual definition but also a clear explanation of its functioning. In its contemporary form, AI operates primarily through algorithms designed to process vast amounts of data, identify patterns, make decisions, and, in many cases, learn from experience (Gentile, 2024: 53-57). One of its core elements is machine learning, a branch of AI that enables computers and machines to imitate human learning processes in order to perform tasks autonomously and improve their performance and accuracy as they are exposed to more data (Baughman *et al.*, 2021). Within this field, deep learning represents a more advanced approach that relies on multilayered artificial neural networks inspired by the human brain; these networks progressively adjust their

internal connections while processing information, which allows them to undertake highly complex tasks such as image recognition or language translation (Gentile, 2024: 53-57). Underpinning both of these techniques are algorithms, logical and mathematical sets of instructions that allow AI systems to analyze data, generate predictions, classify information, group similar content, and refine their outcomes (European Union Agency for Fundamental Rights, 2019: 1-2). Closely connected to these mechanisms is the field of big data, which encompasses the technological developments in data collection, storage, analysis, and application, and is generally characterized by unprecedented increases in the volume, velocity, and variety of data produced, often but not exclusively from online sources (European Union Agency for Fundamental Rights, 2018: 2-3). Finally, an essential component of modern AI is natural language processing (NLP), a specialized branch that allows machines to understand and generate human language, thus enabling applications such as virtual assistants, chatbots, or automatic translation (Jones, 2025). Taken together, these elements demonstrate that artificial intelligence is a dynamic and evolving technology capable of autonomously analyzing data, learning from experience, and making decisions, with the potential to perform increasingly complex tasks that positively impact individuals and society.

II. THE IMPORTANCE OF DATA QUALITY IN ALGORITHMIC DECISION-MAKING

According to the European Union Agency for Fundamental Rights (2018), an algorithm is a set of instructions that enables a computer to transform an input into an output. For instance, it might organize a list of individuals by age. In the field of machine learning, algorithms are designed to make predictions or classifications by analyzing large datasets. Many of these rely on statistical methods —especially regression techniques— to estimate the influence of certain variables on a specific outcome (Baughman *et al.*, 2021). For example, if sufficient data exists, it may be possible to predict an individual's life expectancy based on their alcohol consumption.

It is important to highlight that the development of an algorithm is not limited to the simple automatic execution of rules by a machine: it is a highly human and complex process, involving multiple decisions made by developers, engineers, and managers. These decisions affect everything from the selection and preparation of data to the choice of the statistical model and its final interpretation, (European Union Agency for Fundamental Rights, 2018: 3-4). Therefore, algorithms are not neutral, but rather products of human

processes that directly influence how automated decisions affecting people are made. (Mantelero and Esposito, 2021)

Although algorithms are executed by machines, their design, structure, and results deeply depend on human decisions at all stages of the process: from data collection to model selection and implementation. This human presence is, at the same time, necessary and inevitable, but also a potential source of bias, whether conscious or unconscious (Lendvai and Gosztonyi, 2025:3-4). Algorithms rely on data that can be incomplete, manipulated, biased, or incorrect, which may lead them to reproduce —and even amplify— pre-existing discrimination (García-Marzá, 2023: 101). For this reason, it is essential to approach their development from an ethical perspective that ensures fairness, transparency, and accountability in automated decision-making.

In this regard, data plays a key role in the functioning of AI (Lendvai and Gosztonyi, 2025: 5-6). On the one hand, it is the foundation of AI decision-making, and on the other, depending on its quality, it can lead to biases that result in discrimination. In other words, the quality of the data determines the quality of the algorithm, which is responsible for providing answers, solutions, results during the use of AI.

There are several authors who do not hesitate to state that the use of artificial intelligence gives rise to discriminatory biases, such as the previously mentioned FRA document on discrimination in AI-driven decision-making. Lousada (2024: 101) also affirms that "an AI system is an artifact created, deployed, and used by people. Consequently, the discriminatory biases in AI systems stem from the discriminatory biases that, in general, we have as individuals and as a society." The quality of the data is key to ensuring fair and non-discriminatory decisions. Poor data quality can generate biases that affect the principle of equality (Lousada, 2024: 103 and 116). Below, and based on the work of Lousada (*ibid.*) we will analyze the types of data used by AI, how they are classified, and what ethical and legal risks may arise from their use. In this regard, we can distinguish three main categories:

1. DATA USED BY THE AI SYSTEM

When AI learns from data, it may contain pre-existing biases that are reflected in its decisions. These occur when the data used to train the AI is incomplete, incorrect, or not representative. This causes the system to make incorrect or discriminatory decisions (*ibid.*: 102). Low data quality can result from human intervention in its generation or classification. In the workplace, this is especially common, as seen in employment through digital platforms. (Aragüez, 2022: 8-10) A clear example of this can be seen in platforms like

Uber, where customer-sourced ratings can embed users' biases —often along racial or gender lines— in ways that algorithmic systems then enact, including through automatic driver deactivation at thresholds around a 4.6 average [1] (Levy *et al*., 2016). These systems amplify biased ratings, so that non-White drivers receive approximately 80 % lower ratings and earn 28 % less than their White counterparts (Teng *et al*., 2023). Similarly, field studies in Seattle and Boston show that Black and female riders face higher cancellation rates and longer wait times, highlighting racial and gender biases in service provision —NBER, National Bureau of Economic Research study (Ge *et al*., 2016: 3-18).

2. BIASES DERIVED FROM THE DESIGN OF THE AI SYSTEM ITSELF

Biases can arise not only from how the algorithm processes data but also from how the system is presented and interacted with by users. These design-related biases include both the internal functioning of the algorithm and the structure of the user interface. For instance, interface design can influence how users' access or interpret information. A clear example is a voice assistant that recognizes male voices more accurately than female voices, which can lead to reduced accessibility for women and a user experience shaped by gender bias. Moreover, the default use of female-sounding voices in many virtual assistants has been shown to reinforce harmful gender stereotypes —portraying women as submissive and overly polite— highlighted by the UNESCO and EQUAL skills Coalition Report called *I'd Blush If I Could* (West *et al*: 2019: 9 and 87-88).[2] In addition, an experimental study by Mahmood and Huang (2024: 4-11)[3] found that participants perceived feminine-voiced assistants as warmer and were more likely to interrupt them during errors, whereas gender-ambiguous voices reduced these biases.

---

[1] Uber uses a mechanism that, depending on the city or region, can generate warnings or even automatically deactivate drivers if their average rating falls below a certain threshold (sometimes 4.6, sometimes 4.7). This means that while a score of 4.6 out of 5 would equate to 92% satisfaction and be considered excellent on most scales, Uber's system interprets it as insufficient. Because the system is so demanding, any error can have negative consequences for drivers, not to mention the biases and prejudices present in user ratings.

[2] A UNESCO report showed that voice assistants like Siri, Alexa, and Cortana, by default using female voices that respond with a submissive or apologetic tone, reinforce gender stereotypes and perpetuate the perception of women as "available servants".

[3] An experimental study shows how the perception of assistants' gender (female, male, or ambiguous) impacts user interaction. It is found that female assistants tend to appear warmer after apologies, and ambiguous voices can reduce bias.

3. HUMAN BIASES

These biases do not stem directly from the data or the AI's design but rather arise from human behaviors that influence how the system is used or interpreted. Specifically, biases derived from human actions occur when users, developers, or supervisors of AI unintentionally or deliberately introduce prejudice into its functioning; for example, developers with ideological biases may program a content moderation AI on social media to disproportionately censor certain topics (Lousada, 2024: 112). Additionally, there are risks of invisibility, where AI systems reinforce social inequalities by giving greater visibility to certain groups while marginalizing others. For instance, a search algorithm that prioritizes information in English may render knowledge or cultural perspectives in other languages invisible. As Helm *et al.* (2023: 8-13) demonstrate, "techno-linguistic bias" in AI language technologies systematically excludes underresourced languages and perpetuates epistemic injustice by limiting representation of marginalized linguistic communities.

III. EQUALITY AND NON-DISCRIMINATION AS FUNDAMENTAL RIGHTS IN THE AI CONTEXT

1. DIRECT AND INDIRECT DISCRIMINATION IN THE CONTEXT OF AI

The right to non-discrimination constitutes a fundamental guarantee in the application of artificial intelligence. As highlighted in the previous chapter, the improper use of data and algorithms risks generating decisions tainted by discriminatory biases, with direct consequences for individuals' daily lives. Within the European Union, both direct and indirect discrimination are expressly prohibited under Union law, particularly through anti-discrimination directives such as Directive 2000/43/EC on racial equality and Directive 2000/78/EC on equality in employment.

Direct discrimination, as clarified in *A. v. B.* (Judgment of 11 November 2019, Case C-177/18, ECLI:EU:C:2020:26), occurs when a person is treated less favourably than another in a comparable situation due to a protected characteristic, such as sex, race or ethnic origin, religion or belief, disability, age, or sexual orientation. This form of discrimination is explicit and objective, requiring no proof of discriminatory intent, but only evidence of unjustified unequal treatment.

Indirect discrimination, by contrast, as defined in *Bartsch v. Bosch* (Judgment of 16 April 2008, Case C-427/06, ECLI:EU:C:2008:517), arises where an apparently neutral provision, criterion, or practice places members

of a protected group at a particular disadvantage in comparison with others. While such measures may be justified, this is only permissible if they pursue a legitimate objective and employ means that are both appropriate and necessary in accordance with the principle of proportionality. The Court of Justice of the European Union has repeatedly affirmed that equal treatment requires not merely the absence of discriminatory intent but also the elimination of discriminatory effects in practice, as made explicit in *CHEZ Razpredelenie Bulgaria* (Case C-83/14, Judgment of 16 July 2015).

Beyond case law, this jurisprudence is embedded in a broader legal framework. Article 19 TFEU empowers the Council, acting unanimously on a proposal from the Commission, to adopt measures addressing discrimination based on sex, racial or ethnic origin, religion or belief, disability, age, or sexual orientation, forming the foundation for sectoral legislation in fields such as employment, education, healthcare, and access to goods and services. This is complemented by the Charter of Fundamental Rights, which, since acquiring binding force with the Lisbon Treaty, has reinforced the Union's commitment to equality and non-discrimination. In particular, Article 21 of the Charter prohibits discrimination on an even wider range of grounds —including social origin, genetic features, language, political opinion, membership of a national minority, and property— and applies both to EU institutions and to Member States when implementing Union law. At the regional level, the European Convention on Human Rights adds further protection: Article 14 ensures that Convention rights are secured without discrimination, while Protocol No. 12 establishes a general prohibition of discrimination, thereby extending the protective scope beyond Convention rights.

Notwithstanding this robust legal framework, emerging technological developments —particularly those linked to algorithmic decision-making— pose significant challenges for enforcement. Among these, indirect discrimination proves especially difficult to address in the context of AI, where biases often manifest in systemic and opaque ways. As Zuiderveen Borgesius (2019: 409–411) has noted, algorithmic bias typically operates subtly, complicating its detection, proof, and legal redress through traditional mechanisms. This reality underscores the pressing need to reconsider whether the classic categories of direct and indirect discrimination remain sufficient in the digital age, and to explore possible adaptations of EU anti-discrimination law so as to safeguard fundamental rights in environments increasingly shaped by automated decision-making.

To better illustrate the legal challenges discussed above, particularly the distinction between direct and indirect discrimination in algorithmic decision-making, the following two examples demonstrate how artificial intelligence systems may generate discriminatory outcomes —either explicitly or through

seemingly neutral criteria that disproportionately affect certain protected groups.

An example of direct discrimination in the use of AI: A recruitment company uses an AI system to filter resumes and select candidates for interviews. The algorithm is designed to analyze keywords in the resumes and rate candidates based on their work experience and specific skills. Due to its programming or the historical data used to train it, the AI system systematically favors male candidates in traditionally male-dominated sectors like engineering, while disqualifying or ranking women with the same qualifications lower. This happens because the system interprets that men are more likely to be suitable for the role, as historically more men have held similar positions. This is a clear case of direct discrimination based on sex, as the AI system treats women less favorably simply because they are women, without any objective or proportional justification for the differential treatment.

An example of indirect discrimination: An AI credit scoring system is used to evaluate the creditworthiness of loan applicants, based on various factors such as credit history, employment, and geographic location. The algorithm prioritizes applicants who live in high-income areas, without considering other factors that may affect an individual's ability to repay. Although the AI system does not explicitly discriminate based on race, it ends up disadvantaging racial minorities who live in low-income areas. People from lower-income communities are less likely to receive high scores in the system, even though their economic situation may not fairly reflect their repayment ability. This happens because geographic location and historical income in the area disproportionately influence the evaluation. This is an example of indirect discrimination based on racial origin or ethnic group. While people are not directly discriminated against because of their race, the AI system has a disproportionate impact on a protected group (people from low-income communities, often minorities), creating an indirect disadvantage.

2. IS IT NECESSARY TO REDEFINE THE CONCEPT OF DISCRIMINATION IN THE DIGITAL AGE?

To protect citizens from potential discrimination caused by artificial intelligence, it is essential to ensure that the legal framework in place is capable of safeguarding users' fundamental rights. A central point of debate, as highlighted by Lousada (2024: 117–120), concerns whether algorithmic discrimination can be effectively addressed through existing legal concepts —such as direct or indirect discrimination— or whether the emergence of AI-driven decision-making requires the development of a new legal category: algorithmic discrimination. This discussion has sparked divergent positions in

the academic literature. While some authors argue that the current anti-discrimination laws are sufficient, provided they are properly interpreted and enforced, others contend that the unique features of algorithmic systems —such as opacity, lack of intentionality, and statistical correlation— demand a new normative framework. The following section will explore these differing viewpoints in greater depth, presenting the main arguments from both sides of the debate.

One strand of the academic literature holds that existing legal frameworks are sufficiently robust to address the challenges posed by algorithmic discrimination, without the need to develop a new legal category. Preciado (2021: 6) is one of the most prominent voices in this camp, arguing that algorithmic discrimination represents old forms of discrimination under new appearances and that the legal system already contains the necessary tools to respond to them. According to this view, the use of artificial intelligence does not alter the fundamental nature of discrimination, but merely the context in which it arises. Therefore, the focus should be on the correct application and interpretation of existing norms —particularly those related to indirect discrimination— rather than the creation of entirely new legal definitions.

Similarly, Malgieri (2019; 3-4 and 22-23) takes a nuanced position that emphasizes the adaptability of current EU legal instruments, such as the General Data Protection Regulation[4] (GDPR) and the EU Charter of Fundamental Rights, to situations involving algorithmic decision-making. While acknowledging the complexity and opacity of automated systems, Malgieri contends that the legal notion of discrimination is technically and conceptually capable of encompassing algorithmic harms, provided that existing doctrines are applied flexibly and with sensitivity to the technological context. He argues that the lack of transparency or explainability in AI does not necessarily warrant a new concept, but rather calls for enhanced procedural safeguards and accountability mechanisms within the current legal framework.

Together, these authors suggest that the perceived novelty of algorithmic discrimination may be overstated, and that the core legal principles governing equality and non-discrimination remain applicable —even in the face of technological transformation. In this sense, the challenge lies less in redefining discrimination, and more in ensuring effective enforcement and evidentiary adaptation within the structures already in place.

[4] 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), OJ L 119, 4.5.2016, p. 1–88.

In contrast, several scholars argue that the existing legal framework is insufficient to adequately address the unique characteristics of algorithmic discrimination, and therefore advocate for the development of a new legal concept tailored to these technological realities. Among the most influential voices in this debate are Sandra Wachter, Brent Mittelstadt, and Chris Russell, who in their work *Why fairness cannot be automated: Bridging the gap between EU Non-Discrimination Law and AI* argue that algorithmic discrimination is fundamentally "more abstract, subtle, and intangible" than traditional human forms of discrimination.

According to these authors, algorithms can produce discriminatory outcomes through complex statistical patterns and correlations that do not always align with legally protected categories (Wachter *et al.*, 2020: 11-12). As a result, discriminatory effects may occur without clear intent, and victims may not even be aware they have been disadvantaged. This structural opacity and lack of transparency create significant challenges for detection, investigation, and legal redress, making the classical legal frameworks —based on identifiable actors, direct causality, and intent— less effective in these cases.

To address these gaps, Wachter and colleagues propose incorporating statistical fairness metrics into the legal analysis of discrimination. One such proposal is the use of Conditional Demographic Disparity (CDD) —a quantitative tool designed to evaluate whether and to what extent certain groups are disadvantaged by algorithmic systems. Introduced in the later sections of their study (Wachter *et al.*, 2020: 54, 57, 62), CDD allows courts and regulators to assess disparities in treatment between demographic groups, while still respecting the contextual and flexible nature of EU non-discrimination law. This approach does not seek to replace legal reasoning with statistical metrics, but rather to provide an evidentiary bridge between technical outputs and legal concepts.

Their broader conclusion is that without adapting the legal framework to accommodate these new forms of discrimination, many harms caused by AI systems will remain legally invisible. Thus, the authors call for a rethinking of the current legal doctrine to develop a concept of algorithmic discrimination that is compatible with the structure and logic of automated decision-making.

Occupying a more nuanced position in the ongoing debate, Lousada (2024: 117-120) suggests that while the current anti-discrimination framework —particularly the concepts of direct and indirect discrimination— can be extended to cover algorithmic harms, significant adjustments and interpretative refinements are needed to ensure its effectiveness in the digital context. Rather than proposing a wholly new legal category, Lousada advocates for a series of reforms that bridge the gap between traditional legal norms and the challenges posed by AI-driven decision-making.

First, Lousada emphasizes the importance of adapting existing legal concepts. Although direct and indirect discrimination remain valid categories, their interpretation must evolve to account for discrimination originating not from human intent, but from biased datasets, opaque algorithms, or discriminatory interface designs. This requires courts and regulators to develop a more technically informed understanding of how discrimination may manifest in algorithmic systems.

Second, Lousada highlights the critical role of transparency. The current difficulty in accessing algorithmic logic, source code, or data flows severely impedes the ability of victims to prove discrimination. In this regard, he proposes that non-compliance with transparency obligations —particularly those set out in the forthcoming European Artificial Intelligence Act (AI Act)— could itself constitute a form of discriminatory harm, insofar as it obstructs legal remedies.

Third, the author proposes greater flexibility in evidentiary standards, particularly for cases of indirect discrimination. This could include modifying the traditional test to accommodate the specificities of algorithmic decision-making, or even establishing a quasi-objective presumption of discrimination when a protected group is systematically disadvantaged by an automated system.

Fourth, Lousada argues for a harmonized interpretation between anti-discrimination law and the AI Act. In this view, the preventive mechanisms and risk classifications contained in the AI Act should be seen as complementary to existing non-discrimination norms, creating a layered model of protection that combines ex ante regulation with ex post legal enforcement.

Finally, while he does not definitively call for a new legal concept, Lousada remains open to the potential development of new categories, such as discrimination by omission or refined exceptions based on good faith. These innovations, he suggests, could be explored in future jurisprudence or legislative reform, particularly as courts begin to grapple with the practical consequences of AI deployment in sensitive areas such as employment, credit, and public services.

In sum, the academic debate reflects a tension between continuity and innovation: while some scholars argue that existing anti-discrimination law is sufficiently adaptable to capture algorithmic harms, others maintain that the distinctive features of AI demand a new legal category of algorithmic discrimination. A middle ground, exemplified by Lousada, points to the need for interpretative refinements and procedural reforms within the current framework, in close coordination with the forthcoming AI Act. Ultimately, the challenge lies in ensuring that fundamental rights remain effectively protected in the face of rapidly evolving technological realities.

IV. STRATEGIES FOR THE PROTECTION OF FUNDAMENTAL RIGHTS

1. TRACEABILITY AND EXPLAINABILITY OF ALGORITHMS

In the digital age, algorithmic systems and artificial intelligence play an increasingly significant role in decision-making that affects people's lives. In this context, the principles of transparency and explainability have become essential for ensuring the protection of human rights. In this new chapter, we will analyze them in depth, observing why they are so crucial for protecting users' human rights and also for enabling the challenge of decisions made through AI.

Traceability: Traceability refers to the ability to chronologically interrelate uniquely identifiable entities in a way that is verifiable. (National Institute of Standards and Technology, 2023: 15-16). Therefore, traceability refers to the ability to understand how an automated system works, what data it uses, and under which criteria it makes decisions. It involves the ability to track the entire process that the AI system follows to reach a decision. This includes knowing what data was used, how the model was trained, which algorithms were applied, and what changes were made to the system (Baatout, 2023). As UNESCO points out in its *Recommendation on the Ethics of Artificial Intelligence* (2021:12), "people should be clearly and understandably informed when a decision affecting them has been made by an automated system".

When this traceability is absent, we face what is known as "algorithmic opacity." As Felzmann *et al*. (2020: 3339) point out, this opacity involves a lack of transparency caused by the existence of a "black box," i.e., systems that lack explanatory capability and whose internal logic is inaccessible or incomprehensible to users and, in many cases, even to the developers themselves. It is not only that we do not know what decision is made, but that we are completely unaware of how and why.

This lack of traceability creates worrying situations, especially when algorithms make decisions that directly affect people's rights and opportunities. For example, when a bank loan, social aid, or job position is denied, and the algorithm does not allow us to understand the criteria used or offer the possibility to challenge the decision, it becomes a form of an unappealable verdict, potentially based on partial, erroneous, or biased data, with no right for the affected person to an explanation (Blázquez, 2022: 268).

Beyond the individual case, the greater risk, according to Blázquez Ruiz (*ibid.*), is the consolidation of a hyper-technician society, where even the elites do not fully understand the functioning of the tools that govern key decisions. This scenario fosters structural inequality and a loss of democratic and citizen control over the systems that increasingly mediate our daily lives (*ibid.*: 268). For all these reasons, traceability is not just a technical issue, but an essential

element to guarantee accountability, justice, and respect for fundamental rights in digital environments

Explainability: means ensuring that algorithmic decisions, as well as any data driving those decisions, can be explained to end-users and other stakeholders in non-technical terms. This includes information that allows an explanation of the general functioning of the system, the specific use of data within the system, and individual decisions taken by the system. (Felzmann *et al.*, 2020: 3347)

In the current context of the rapid development of artificial intelligence, explainability emerges as a fundamental condition to ensure the ethical and just application of automated systems, particularly those based on deep learning. This requirement has been emphasized by authors such as López de Mántaras (2021), who argues that algorithms must incorporate explanatory modules that allow understanding and interpreting the decisions they generate. Explainability should not be conceived as an optional addition, but as an indispensable feature of any intelligent system, especially when its decisions directly impact individuals' social or legal rights (Turri, 2022).

This concern has also been reflected at the institutional level. The European Commission, through the High-Level Expert Group on Artificial Intelligence, included explainability as one of the key *Principles in its guidelines for trustworthy AI* (2019: 9). According to this body, AI systems must be explainable so that users can understand and, if necessary, challenge the results, especially when these affect fundamental rights. In a constantly evolving social and technological context, where artificial intelligence systems are actively involved in decisions that affect fundamental rights, explainability emerges as an irreplaceable principle for the effective protection of human rights.

Traceability and explainability are not simply technical features, but essential guarantees of transparency, accountability, and respect for fundamental rights. Both academic voices and institutional initiatives —such as the European Commission's *Guidelines for trustworthy AI* (2019)— stress their indispensable role in enabling individuals to understand, question, and contest algorithmic decisions. Advancing toward more transparent and explainable AI is therefore not only a technical challenge, but also a democratic imperative: without it, artificial intelligence risks consolidating opaque power structures that undermine justice, equality, and citizen control.

2. PREVENTIVE APPROACHES: METHODS TO AVOID BIAS BEFORE IMPLEMENTING AI

Before an artificial intelligence system is put into operation, it is essential to establish mechanisms that ensure its impartiality and respect for funda-

mental rights. Preventing discriminatory biases is not only a technical issue but also an ethical and legal one, especially when AI is applied in sensitive areas such as employment, education, healthcare, or justice. Many of the *a priori* prevention mechanisms aimed at avoiding discrimination produced by AI are found in the European Union Agency for Fundamental Rights (FRA) document on data quality and artificial intelligence (2019: 2-4, 13).

To prevent discrimination in AI systems, the first crucial aspect, according to the FRA, is data quality. While large volumes of data may give the impression of accuracy, this is misleading if the quality of the data is not assessed. Statistical accuracy and the ability to reflect the real world depend not only on the amount of data but also on its quality. A key element of data quality is whether the data are fit for purpose —meaning they are suitable for the specific task they are intended to support. Evaluating whether data serve their intended purpose helps to identify errors and potential risks in data-driven systems.

According to the FRA, data must be of high quality to prevent discriminatory bias in AI systems. This means data should be accurate, complete, consistent, up-to-date, valid, non-duplicated, available, and traceable to a reliable source (provenance). These characteristics ensure that the data are fit for their intended purpose and reduce the risk of bias during model training and decision-making.

Another important concept introduced is measurement error, which refers to the extent to which the data accurately represent what they are supposed to measure. For example: Is income a reliable indicator of creditworthiness? What defines a "good employee"? Being punctual? Performing well? Understanding these dimensions of data quality and measurement is essential to identify and mitigate potential sources of bias, especially in training data used in machine learning models.

Another key aspect of data quality, according to the FRA, is labeling. Human labeling of outcome data (e.g., assigning categories to images) is essential for evaluating dataset bias and algorithm performance, but it can introduce measurement errors if quality control is lacking. Poor or biased labeling leads to unfair and inaccurate models.

The FRA warns of the risk of representation error, which arises when training data fail to reflect the target population or phenomenon; for instance, using nationality as a proxy for origin may exclude naturalized individuals and distort results. Such shortcomings directly undermine two core pillars of data quality: reliability, understood as consistency across time and conditions, and validity, referring to whether data genuinely measure the phenomenon they are intended to capture.

In machine learning, high data volume alone does not guarantee quality. If data lack validity or representativeness, bias is not eliminated but automated

at scale. The FRA warns that large datasets can amplify systematic errors, creating a false sense of accuracy. Understanding and addressing these risks is essential for responsible AI development.

Another highly relevant aspect that deserves to be highlighted in this work is the HUDERIA (Human Rights, Democracy and the Rule of Law Impact Assessment) document (Conseil of Europe, 2024), which establishes a methodology for evaluating the risks and impacts of artificial intelligence (AI) systems from the perspective of human rights, democracy, and the rule of law. This document was adopted by the Artificial Intelligence Committee (CAI) of the Council of Europe in November 2024 and builds on the previous work of CAHAI and the Alan Turing Institute.

HUDERIA establishes a preventive risk mechanism designed to anticipate and manage the potential negative effects of AI use, structured in four phases. The first phase, COBRA, helps understand the system's context and map potential risks throughout its lifecycle. The second phase, SEP, ensures that the voices of those who could be affected are heard by engaging stakeholders to understand risks from their perspective. The third stage, RIA, conducts a detailed assessment of risks and impacts, focusing particularly on the protection of democracy, the rule of law, and human rights. Finally, the MP phase defines strategies to mitigate identified risks and monitor the system's operation throughout its lifecycle, functioning as a structured risk audit. We will not delve further into this topic, though its importance and relevance are undeniable. For more detailed information, the following link provides an expanded explanation.[5]

To prevent discrimination caused by AI, the report of European Union Agency for Fundamental Rights called Getting the Future Right: Artificial Intelligent and Fundamental Rights (2020: 7, 10, 64, 66, 72, 96) emphasizes the importance of conducting impact assessments that go beyond technical performance and include a fundamental rights perspective. While traditional assessments often focus on factors like accuracy or cybersecurity, this is not enough if the social and human effects of AI systems are ignored. Impact assessments serve as preventive tools, helping organizations and governments identify potential risks to rights such as privacy and non-discrimination. Although some legal frameworks (like the General Data Protection Regulation) already require such evaluations, they often overlook whether AI systems may infringe on human rights. We cannot forget another fundamental tool, the HUDERIA document, which we consider contains excellent recommendations and practical guidelines to ensure the safety in the operation of an AI

[5] https://is.gd/WS2R0D

system. For AI to be used safely and ethically, it is essential to ensure effective oversight and to establish clear responsibilities for those who design, deploy, and regulate these systems.

Prevention of Discriminatory Biases in the EU Regulation on Artificial Intelligence

Regulation (EU) 2024/1689 represents a major step forward in the regulation of artificial intelligence within the European Union, with a particular focus on preventing discriminatory bias, especially in high-risk systems. To this end, the regulation sets out a series of technical, organizational, and ethical requirements aimed at protecting fundamental rights.

Among the most relevant provisions, Article 9 requires the implementation of a risk management system capable of identifying, analyzing, and assessing potential impacts on people's health, safety, and rights, including the risk of discrimination, as well as adopting preventive measures to reduce the likelihood of biased outcomes. Complementarily, Article 12 strengthens transparency by mandating the maintenance of accurate records on system operation, enabling audits, external oversight, and retrospective detection of possible biases. This traceability is combined with human oversight, regulated in Article 14, which ensures the possibility of intervention to correct errors and prevent discriminatory decisions, thus consolidating the role of human judgment as a safeguard against unfair automation.

From a technical perspective, Article 15 establishes criteria for accuracy, robustness, and cybersecurity that reduce the susceptibility of systems to systematic errors that could generate bias, while Article 17 obliges AI providers to implement a quality management system ensuring that the design, development, and use of data comply with legal and technical standards, avoiding discriminatory outcomes derived from biased or mislabeled information. Article 27 reinforces this preventive approach by requiring risk assessments prior to the deployment of any system, explicitly considering discriminatory impacts on vulnerable groups and analyzing how AI might affect individuals based on gender, race, religion, or disability.

The supervision of authorities, regulated in Article 53, ensures that AI applications do not create inequalities and that decisions are verifiable, while Article 67 strengthens transparency by requiring the public disclosure of the methods and data used, enabling the detection and correction of bias from both the design and operational phases of systems. Finally, Article 70 establishes the need for periodic reviews of preventive measures to assess their effectiveness, ensuring that AI systems continue to uphold non-discrimination principles over time.

Taken together, the regulation emphasizes a comprehensive approach that combines technical, ethical, and legal measures —including fundamental rights impact assessments, bias detection testing before deployment, and rigorous data quality management— with the aim of mitigating the reproduction or amplification of existing inequalities. Thus, Regulation (EU) 2024/1689, along with reports such as those from the FRA, provides a solid framework for the responsible development and oversight of AI in Europe, promoting justice, fairness, and the protection of human rights.

3. CORRECTIVE APPROACHES: EX POST SOLUTIONS TO MITIGATE NEGATIVE IMPACTS

Even with strong preventive measures, harmful impacts from AI may still occur. This section addresses corrective (ex post) approaches, focusing on how to respond when fundamental rights are violated. It highlights the need for effective systems of accountability, reparation, and access to legal or administrative remedies, along with the importance of retroactive transparency and the roles of both public and private actors in ensuring restitution.

According to Arenós (2025), specific legal frameworks are needed to assign responsibility for harm caused by AI systems, allowing victims access to judicial or administrative remedies. The creation of civil liability insurance for AI-related damages is also proposed to encourage safer and more ethical AI practices.

The European Union Agency for Fundamental Rights (FRA, 2020: 13, 67 and 81) highlights that effective access to justice for AI decisions is essential to protect fundamental rights. Citizens must be informed about AI use, understand its workings, and know how to file complaints. Access to remedies is not only a right in itself, as stated in Article 47 of the Charter of Fundamental Rights of the EU (2000), but also crucial to exercising other rights, like the right to a fair trial. States must ensure access to judicial or alternative routes for reparation.

Challenges such as AI's technical complexity and intellectual property constraints hinder transparency. Some entities adopt simpler AI models in sensitive areas and promote accountability and citizen oversight initiatives.

Furthermore, Mantelero and Esposito (2021:55-56) stress the importance of conducting human rights impact assessments before and after AI deployment. These assessments, performed by independent bodies, help identify and mitigate risks while ensuring compliance with ethical and legal standards.

In 2025, Meritxell Borràs, president of the Catalan Data Protection Authority (APDCAT), introduced an innovative methodology for supervising

and humanizing artificial intelligence, aimed at assessing its impact on fundamental rights such as privacy, equality, and non-discrimination (Autoridad Catalana de Protección de Datos, 2025a, 2025b). The proposal is based on a risk matrix that evaluates both the probability of rights violations occurring and the severity of their potential consequences, depending on the characteristics of the AI system and its context of use. This approach is designed to help identify high-risk applications —such as those used in justice, security, or healthcare— before harm occurs, allowing public and private decision-makers to proactively adapt systems to prevent violations. Beyond identifying risks, the methodology promotes the adoption of corrective measures through continuous review and auditing protocols, ensuring compliance with ethical and legal standards. Overall, this initiative represents a significant step toward establishing more transparent and robust regulatory frameworks that guarantee human oversight and protect citizens' rights in the development and use of AI.

Additionally, Aritz Obregón Fernández and Guillermo Lazcoz Moratinos (2021: 2-3, and 20-27) highlight the critical need for meaningful human control (MHC) over high-risk AI systems, especially in sensitive areas like justice, public safety, and health. MHC requires human supervision at all stages of AI decision-making to protect fundamental rights, ensuring that AI outputs can be reviewed, validated, or modified when necessary. They argue this principle must be explicitly integrated into the design, development, and deployment of AI, supported by both International Law and European Union law.

The authors discuss how EU regulations, such as the 2021 proposed AI Regulation, adopt a risk-based approach demanding human oversight for high-risk AI but call for clearer emphasis on MHC due to AI's disruptive potential. They also note practical challenges, particularly the need to train and certify human operators with technical skills to effectively supervise complex AI systems and intervene appropriately.

Corrective Approaches a Posteriori in the EU Artificial Intelligence Regulation

The Artificial Intelligence Regulation establishes crucial corrective mechanisms to address negative impacts of AI when preventive measures are insufficient. Article 65 provides for the creation of the European Artificial Intelligence Board, a key body for coordinated oversight at the European level. While it does not have direct sanctioning powers, its main role is to promote technical coherence and cooperation among national authorities, supporting joint actions to address systemic or compliance issues and facilitate the adoption of effective corrective measures against detected biases or violations.

Once discriminatory systems are identified, authorities must act immediately to rectify the situation and prevent infringements of fundamental rights. Article 68 recognizes the right of individuals affected by discriminatory automated decisions to obtain redress, allowing them to appeal to competent authorities and receive compensation, thereby reinforcing accountability of system operators. Chapter IX, Section 4, establishes users' rights against violations by AI systems, ensuring clear mechanisms to challenge decisions and obtain explanations for high-risk automated processes. Articles 85–87 cover the right to submit complaints to market surveillance authorities, receive detailed explanations of individual automated decisions, and provide protection for whistleblowers in line with Directive (EU) 2019/1937.

Section 5 of Chapter IX focuses on corrective measures related to mitigating risks from general-purpose AI models, including mechanisms for supervision, compliance, and monitoring of providers. Article 90 allows expert groups to issue alerts on models presenting systemic risks, enabling the Commission and the AI Office to take corrective actions, such as marketing restrictions, withdrawal, or recall of models if significant risks are detected. Articles 91 and 92 empower the Commission to request information from providers and perform technical assessments, including access to source code, to verify compliance with safety and fundamental rights requirements.

Overall, these corrective measures are crucial for managing AI risks and ensuring accountability, oversight, and transparency. By combining the European AI Board, impact assessments, human supervision, and targeted actions, the Regulation promotes both proactive and reactive protection of citizens' fundamental rights in a complex, automated environment.

V. FUTURE PERSPECTIVES OF EUROPEAN LAW IN RELATION TO AI AND DISCRIMINATION

In this section, we will examine some cases from the jurisprudence of the Court of Justice of the European Union. Although the CJEU has not yet issued any rulings directly related to autonomous artificial intelligence systems, it has begun to address issues concerning the automated use of personal data and the effects these systems may have on fundamental rights, particularly the right to non-discrimination and data protection

1. CASE C-311/18, *SCHREMS II*

The C-311/18 case, known as *Schrems II*, represents a landmark in the evolution of European data protection law and the regulation of international

transfers of personal information (Mildebrath, 2020). The dispute arose from the transfer of personal data by Facebook Ireland Ltd. to its parent company in the United States, carried out under the Privacy Shield framework and the Standard Contractual Clauses (SCCs) approved by the European Commission (EPRS, 2020). Austrian lawyer and privacy activist Maximilian Schrems challenged the legality of these transfers before the Irish Data Protection Commission, arguing that U.S. law did not provide adequate safeguards against the extensive access of personal data by agencies such as the NSA, thereby breaching the standards of the GDPR and the rights enshrined in the Charter of Fundamental Rights of the European Union, particularly those concerning privacy (Art. 7), data protection (Art. 8), and effective judicial remedy (Art. 47).

In its judgment of 16 July 2020, the Court of Justice of the European Union invalidated the Privacy Shield, concluding that the U.S. regime failed to provide a level of protection essentially equivalent to that required by the EU and lacked effective judicial redress mechanisms. By contrast, the SCCs were upheld as valid, but subject to strict conditions: their use requires that the recipient country can ensure —either through domestic legislation or supplementary measures adopted by the parties— a level of protection essentially equivalent to the GDPR. This entails a prior assessment of the legal framework and suspension of transfers when adequate safeguards cannot be guaranteed.

Although *Schrems II* does not directly address artificial intelligence, its reasoning on data protection and fundamental rights is highly relevant for the governance of automated decision-making. The judgment highlights the necessity of transparency, oversight, and effective redress —principles that, as noted by Costello (2020), are central to ensuring that international data transfers comply with EU standards of privacy and security. While Costello's analysis focuses on cross-border data flows, these same safeguards can be extrapolated to the context of opaque or "black box" algorithms capable of making decisions with significant social and legal consequences. Moreover, the principle of equivalence in international transfers established by the judgment implies that any personal data processing —including that conducted through AI— must meet standards comparable to those of the GDPR, thereby preventing bias or discriminatory practices. In this way, Schrems II reinforces accountability, responsibility, and the protection of fundamental rights, providing a normative framework which, although originally conceived for international data transfers, is fully applicable to the oversight of AI's impact in digital society (Judgment of 16 July 2020).

2. CASE C-154/21, *UI V. ÖSTERREICHISCHE POST AG*

The ruling of the Court of Justice of the European Union in *UI v. Österreichische Post AG* (C-154/21) addresses the creation of automated political profiles without the data subject's consent and clarifies the scope of compensation for non-material damages under Article 82 of the GDPR (Court of Justice of the European Union, 2023). The Court established that compensation is not granted automatically in the event of a breach of the Regulation; rather, the claimant must demonstrate (i) the existence of a violation, (ii) actual damage —whether material or non-material— and (iii) a causal link between the violation and the harm suffered. Importantly, the Court emphasized that there is no minimum severity threshold for non-material damages, thereby strengthening the protection of fundamental rights and acknowledging the subjective impact of algorithmic processing (Court of Justice of the European Union, 2023).

From the perspective of artificial intelligence, the case is highly relevant. The processing carried out by Österreichische Post AG relied on automated statistical methods to infer estimated political profiles from demographic and socioeconomic characteristics. Although not formally classified as AI, these practices replicate core dynamics of algorithmic systems, including opacity in decision-making and the prediction of behavior without meaningful human oversight. The judgment illustrates that such automated processing can inflict non-material harm —such as distress, loss of trust, or a sense of surveillance— that is nonetheless legally protected under the GDPR (Rooney & Coll, 2024; Brams *et al*., 2023).

According to Brams *et al*. (2023), organizations must provide data subjects with information on the specific recipients of personal data when requested, except in cases where this is impossible or when the request is manifestly unfounded or excessive. This obligation reinforces transparency and accountability in automated processing: individuals should not only be informed that their data is used in profiling but also be able to identify who receives, or may receive, that information.

From both a legal and ethical standpoint, the ruling strengthens the connection between fundamental rights and the operation of automated systems. It requires organizations to document data flows, ensure human oversight, and recognize that even intangible algorithmic harms can give rise to legal liability. This sets an important precedent for balancing technological innovation with individual rights and transparency in AI governance across Europe.

Although neither *Schrems II* (C-311/18) nor *UI v. Österreichische Post AG* (C-154/21) directly concern autonomous AI systems, both establish

principles highly relevant for their regulation. *Schrems II* underscores the need for equivalence, transparency, and effective redress in data processing, while *UI v. Österreichische Post AG* highlights the risks of automated profiling and confirms liability for even non-material harms. Together, they reinforce a normative framework centered on accountability, human oversight, and the protection of fundamental rights, offering valuable guidance for addressing discrimination and other challenges posed by AI in Europe.

VI. CONCLUSIONS

Artificial Intelligence has rapidly entered our societies, transforming multiple areas of daily life, from healthcare and education to job recruitment and access to public services. However, alongside its benefits and potential, AI also presents serious ethical and legal challenges —among which the risk of discrimination stands out. This study has shown that AI systems, being developed and trained by humans, are not neutral: they reflect, reproduce, and can even amplify existing biases and inequalities in society.

In this context, it has been identified that the main causes of algorithmic discrimination stem from three key factors: poor data quality, implicit biases embedded in the system's technical design, and human prejudices that influence all stages of AI development and implementation. For example, the use of historical data that is biased or incomplete can result in unjust or incorrect decisions that disproportionately affect vulnerable groups. Likewise, interface and algorithm design can unintentionally incorporate stereotypes based on gender, race, or class, replicating existing forms of social and labor exclusion.

One of the most relevant contributions of this work has been the analysis of the European regulatory framework, particularly the Artificial Intelligence Act (Regulation (EU) 2024/1689), which constitutes a landmark in regulating the development and use of AI systems. This regulation introduces measures such as traceability, explainability, human oversight, risk assessments, and accountability mechanisms aimed at protecting fundamental rights in an increasingly automated world. It emphasizes both preventive (ex ante) and corrective (ex post) strategies to reduce the likelihood and consequences of discriminatory automated decisions.

Despite these legal advances, the study also reveals certain limitations of the current framework in addressing the challenges posed by modern algorithmic systems. In particular, it raises the question of whether classical legal concepts of direct and indirect discrimination are sufficient to deal with more subtle and systemic algorithmic harms. Some scholars advocate for the

recognition of a new legal category — "algorithmic discrimination"— that can better capture the harm caused by opaque, unintentionally biased, and technically complex decision-making systems. Others propose adapting and reinterpreting existing concepts and strengthening evidentiary standards, transparency requirements, and participatory safeguards to ensure the continued relevance of anti-discrimination law in this digital context.

Additionally, recent rulings by the Court of Justice of the European Union, such as *Schrems II* and *UI v. Österreichische Post AG*, underscore the need for transparency, accountability, and protection of individual rights in automated systems. These decisions provide a normative framework that reinforces ethical governance, human oversight, and the safeguarding of fundamental rights, offering important guidance for the regulation of AI in Europe.

Therefore, it is concluded that the challenge lies not only in improving the technical performance of algorithms, but also in adopting an interdisciplinary approach that brings together ethical principles, legal protections, and democratic oversight in the use of AI. Algorithmic justice requires political will, effective legislation, responsible technological development, and an informed and empowered civil society that demands transparency, accountability, and the right to redress.

In summary, ensuring fairness and non-discrimination in the age of artificial intelligence is not merely a desirable goal, but an urgent necessity to protect human dignity, prevent the reproduction of historical injustices, and build a more inclusive, just, and rights-respecting digital future.

Bibliography

Aragüez, L. (2022). Desafíos de la digitalización de las relaciones laborales: algoritmos digitales, robotización y trabajo a distancia. *E-Revista Internacional de la Protección Social,* 7 (1), 8-10. Available in: https://dx.doi.org/10.12795/e-RIPS.2022.i01.01.

Arenós Karsten, J. (2025). Closing the gap: Fair victim compensation in the EU AI liability regime. *European Student Think Tank*, 5-5-2025. Available in: https://is.gd/ZAlcz9.

Autoridad Catalana de Protección de Datos (APDCAT) (2025a). *Cataluña presenta un modelo pionero en Europa para desarrollar soluciones de IA respetuosas con los derechos fundamentales*. Available in: https://is.gd/8C9NXV.

Autoridad Catalana de Protección de Datos (APDCAT) (2025b). *La APDCAT promueve la implementación del modelo catalán FRIA en Brasil para garantizar los derechos fundamentales en el diseño y uso de aplicaciones de IA.* Available in: https://is.gd/UBj5L2

Baatout, A. (2023). Why traceability is important in artificial intelligence. *Adesso* [blog], 27-11-2023. Available in: https://is.gd/WjZ7sN.

Baughman, A., Hay, C. y Soule, K. (2021). What is machine learning? *International Business Machine*, 22-9-2021.

Blázquez Ruiz, F. J. (2022). La paradoja de la transparencia en la IA: opacidad y explicabilidad. Atribución de responsabilidad. *Revista Internacional Pensamiento Político*, 17, 261-272. Available in: https://doi.org/10.46661/revintpensampolit.7526.

Brams I., Docherty, C., Smyth, A. and Wybitul, T. (2023). CJEU Sets High Bar for Responses to Data Subject Access Requests. *Global Privacy and Security Compliance Blog* [blog], 5-5-2023. Available in: https://is.gd/e6QzlY.

Costello, R. Á. (2020). *Schrems II: Everything is illuminated? European Papers- European Forum*, 5 (2), 1045-1059. Available in: https://www.europeanpapers.eu/en/europeanforum/schrems-II-everything-is-illuminated

European Union Agency for Fundamental Rights (2018). *Big data: Discrimination in data-supported decision making.* Report of the European Union Agency for Fundamental Rights (Vienna, Austria). Luxembourg: Publications Office of the European Union.

European Union Agency for Fundamental Rights (2020). *Getting the future right. Artificial intelligence and fundamental rights.* Report of the European Union Agency for Fundamental Rights (Vienna, Austria). Luxembourg: Publications Office of the European Union.

Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamo-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics,* 26 (6), 3345-3348. Available in: https://doi.org/10.1007/s11948-020-00276-4.

García-Marzá, D. (2023). Ética digital discursiva: de la explicabilidad a la participación. *Revista Internacional de Filosofía*, 90, 99-114.

Gentile, N. (2024). *Entiende la tecnología: Desde la caída de Megaupload hasta los secretos de la inteligencia artificial*. Barcelona: Editorial B.

Ge, Y., Knittel, C. R., MacKenzie, D. y Zoepf, S. (2016). *Racial and gender discrimination in transportation network companies* (NBER Working Paper No. 22776). Cambridge, MA: National Bureau of Economic Research. Available in: http://www.nber.org/papers/w22776

High-Level Expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI*. Set up by the European Commission. Brussels.

Helm, P., Bella, G., Koch, G, and Giunchiglia, F. (2023). *Diversity and language technology: How technolinguistic bias can cause epistemic injustice*. Available in: https://doi.org/10.48550/arXiv.2307.13714.

Jones, M. T. (2025). A beginner's guide to natural language processing. *IBM Developer,* 23-4-2025. Available in: https://is.gd/QUhUrc.

Lendvai, G. F. and Gosztonyi, G. (2025). Algorithmic Bias as a Core Legal Dilemma in the Age of Artificial Intelligence: Conceptual Basis and the Current State of Regulation. *Laws*, 14 (3), 41. Available in: https://doi.org/10.3390/laws14030041.

Levy, K., Barocas, S., Rosenblat, A. and Hwang, T. (2016). Discriminating Tastes: Customer Ratings as Vehicles for Bias. *OnLabor.* [blog] 22-10-2015. Available in: https://is.gd/0YtW2R.

Lopez de Mantaras, R. (2021). La inteligencia artificial nunca seré como la humana. *La Vanguardia*, 29-3-2021. Available in: https://is.gd/weP9Yj.

Lousada Arochena, J. F. (2024). Inteligencia artificial y sesgos discriminatorios: ¿es necesario un nuevo concepto de discriminación algorítmica? *IgualdadES*, 11, 97-123. Disponible en: https://doi.org/10.18042/cepc/IgdES.11.04.

Mahmood, A. and Huang, C.-M. (2024). Gender biases in error mitigation by voice assistants. *Proceedings of the ACM on Human-Computer Interaction,* 8 (60), 1-27. Available in: https://doi.org/10.1145/3637337.

Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law and Security Review*, 35 (5). Available in: https://doi.org/10.1016/j.clsr.2019.05.002.

Mantelero, A. and Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in data-intensive AI systems development. *Computer Law and Security Review*, 41, 1-57. Available in: https://doi.org/10.1016/j.clsr.2021.105561.

Mildebrath, H. (2020). *The CJEU judgment in the Schrems II case.* European Parliamentary Research Service. Available in: https://is.gd/FfXs1m.

Obregón Fernández, A. y Lazcoz Moratinos, G. (2021). La supervisión humana de los sistemas de inteligencia artificial de alto riesgo: aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea. *Revista Electrónica de Estudios Internacionales*, 42, 1-29. Disponible en: https://doi.org/10.17103/reei.42.08.

Preciado Domènech, C. H. (2021). Algoritmos y discriminación en la relación laboral. Jurisdicción Social. *Revista de la Comisión de lo Social de Juezas y Jueces para la Democracia*, 223, 5-24.

Rooney, C. and Coll, A. (2024). Summary of 2023's key CJEU data protection judgments. *Arthur Cox LLP* [blog], 24-1-2024. Available in: https://is.gd/YHY9m8.

Russell, S. and Norvig, P. (2016). *Artificial Intelligence. A modern Approach (*4th ed.). Madrid: Pearson.

Teng, C., Botelho, T. and Sudhir, K. (2023). Ratings Systems Amplify Racial Bias on Gig Economy Platforms. *Yale School of Management*, 14-8-2023. Available in: https://is.gd/EaPUZO.

Turri, V. (2022). What is Explainable AI? *Carnegie Mellon University, Software Engineering Institute's Insights* [blog], 17-6-2022. Available in: https://is.gd/oqEr97.

UNESCO (2021). *Recommendation on the ethics of artificial intelligence*. Paris: UNESCO.

U.S. National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework.* (AI RMF 1.0). Gaithersburg, MD: National Institute of Standards and Technology. Available in: https://doi.org/10.6028/NIST.AI.100-1.

Wachter, S., Mittelstadt, B. and Russell, C. (2020). Why fairness cannot be automated: Bridging the gap between EU Non-Discrimination Law and AI. *SSRN Electronic Journal.* Available in*:* https://doi.org/10.48550/arXiv.2005.05906. c

West, M., Kraut, R., and Ei Chew, H. (2019). I'd blush if I could: Closing gender divides in digital skills through education. Paris: UNESCO. EQUAL Skills Coalition. Available in: https://doi.org/10.54675/RAPC9356

Zuiderveen, F. (2019). Price discrimination, algorithmic decision-making, and European non-discrimination law. *European Business Law Review*, 31, 401-422. Available in: https://doi.org/10.54648/EULR2020017.

Case law sources

Judgment of 4 May 2023, *UI v. Österreichische Post AG*, Case C-154/21. ECLI:EU:C:2023:370

Judgment of 11 November 2019, *A. v. B.*, Case C-177/18 ECLI:EU:C:2020:26

Judgment of 16 April 2008, *Bartsch v. Bosch*, Case C-427/06. ECLI:EU:C:2008:517

Judgment of 16 July 2015, *CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia*, Case C-83/14. ECLI:EU:C:2015:480

Judgment of 16 July 2020, *Data Protection Commissioner v. Facebook Ireland Ltd and Maximillian Schrems*, Case C 311/18, ECLI:EU:C:2020:559.