

Deusto Journal of Human Rights

Revista Deusto de Derechos Humanos

No. 14/2024

DOI: <https://doi.org/10.18543/djhr142024>

ARTICLES / ARTÍCULOS

Facing fundamental rights in the age of preventive *ex ante* AI: a contemporary form of discrimination

La encrucijada de los derechos fundamentales en la era del control *ex ante* asociado a la IA preventiva: Una nueva forma de discriminación

M^a Teresa García-Berrio Hernández

<https://doi.org/10.18543/djhr.3191>

E-published: December 2024

Copyright (©)

Deusto Journal of Human Rights / Revista Deusto de Derechos Humanos is an Open Access journal; which means that it is free for full and immediate access, reading, search, download, distribution, and reuse in any medium only for non-commercial purposes and in accordance with any applicable copyright legislation, without prior permission from the copyright holder (University of Deusto) or the author; provided the original work and publication source are properly cited (Issue number, year, pages and DOI if applicable) and any changes to the original are clearly indicated. Any other use of its content in any medium or format, now known or developed in the future, requires prior written permission of the copyright holder.

Derechos de autoría (©)

Deusto Journal of Human Rights / Revista Deusto de Derechos Humanos es una revista de Acceso Abierto; lo que significa que es de libre acceso en su integridad inmediatamente después de la publicación de cada número. Se permite su lectura, la búsqueda, descarga, distribución y reutilización en cualquier tipo de soporte sólo para fines no comerciales y según lo previsto por la ley; sin la previa autorización de la Editorial (Universidad de Deusto) o la persona autora, siempre que la obra original sea debidamente citada (número, año, páginas y DOI si procede) y cualquier cambio en el original esté claramente indicado. Cualquier otro uso de su contenido en cualquier medio o formato, ahora conocido o desarrollado en el futuro, requiere el permiso previo por escrito de la persona titular de los derechos de autoría.


Deusto Journal of Human Rights

ISSN: 2530-4275 • ISSN-e: 2603-6002, No. 14/2024, Bilbao

© Universidad de Deusto • <http://djhr.revistas.deusto.es/>

Facing fundamental rights in the age of preventive *ex ante* AI: a contemporary form of discrimination

La encrucijada de los derechos fundamentales en la era del control *ex ante* asociado a la IA preventiva:
Una nueva forma de discriminación

M^a Teresa García-Berrio Hernández 

Universidad Complutense de Madrid. España

teresag-berrio@der.ucm.es

ORCID: <https://orcid.org/0000-0002-4205-4184>

<https://doi.org/10.18543/djhr.3191>

Submission date: 06.06.2024

Approval date: 22.11.2024

E-published: December 2024

Citation / Cómo citar: García-Berrio, M^a Teresa. 2024. «Facing fundamental rights in the age of preventing *ex ante* AI: a contemporary form of discrimination.» *Deusto Journal of Human Rights*, n. 14: 101-125. <https://doi.org/10.18543/djhr.3191>

Summary: 1. Artificial Intelligence and human condition: for an ethical use of AI. 2. EU Artificial Intelligence Act: a new roadmap on fundamental rights risk management in the face of AI. 3. Preventive risk control: a contemporary form of discrimination. 4. Principle of Non-maleficence: Prevention of harm and preservation of human dignity in the face of the risk of AI. 5. Mitigating the discriminatory impact of biases in AI algorithms: seeking the beneficence principle. Conclusions. References.

Abstract: As Artificial Intelligence (AI) systems become increasingly integrated into the social fabric of contemporary communities, ethical considerations surrounding their impact on fundamental rights have come to the fore. Indeed, the growing significance of AI has recently prompted a pivotal discourse within academic and policy circles in Europe concerning the development of an ethical framework for human-centric AI. As part of a broader research project examining the implications of AI on fundamental rights, particularly the right to non-discrimination, our objective is to present a preliminary overview of fundamental rights' risk management in the context of AI. In light of the significant impact of AI on vulnerable individuals and minorities, our discussion will subsequently address critical areas of concern related to the EU AI Act, including algorithmic bias and its constituent elements of discrimination based on ethnicity or religion.

Keywords: AI, ethics, fundamental rights, algorithmic biases, EU policies

Resumen: A medida que los sistemas de Inteligencia Artificial se integran cada vez más en el tejido social de las comunidades contemporáneas, las consideraciones éticas en torno a su impacto sobre los derechos fundamentales cobran más fuerza. En este sentido, tanto en círculos académicos como políticos europeos se ha propagado en los últimos años un debate recurrente sobre la viabilidad de construir un marco ético favorable a una dimensión antropocéntrica de la IA. Como parte de un proyecto de investigación más amplio que examina las implicaciones de la IA sobre los derechos fundamentales, en particular el derecho a la no discriminación, nuestro objetivo es presentar una visión preliminar de la gestión del riesgo de los derechos fundamentales en el contexto de la IA. A la luz del significativo impacto de la IA sobre las personas vulnerables y las minorías, nuestro estudio abordará asimismo aquellas áreas críticas relacionadas con el Reglamento europeo de la IA, incluido el sesgo algorítmico y sus elementos constitutivos de discriminación por motivos étnicos o religiosos.

Palabras clave: Inteligencia artificial, ética, derechos fundamentales, sesgos algorítmicos, políticas europeas.

1. Artificial Intelligence and human condition: for an ethical use of AI¹

The implementation of Artificial Intelligence (AI) in our daily-basis routines represents an unprecedented anthropological disruption, with a direct impact on all natural, economic, and social structures of human communities. The advent of recent technological advances has posed a challenge to our ethical consciousness, particularly in light of the illusion of postmodern individual autonomy. This illusion is strongly marked by the high price paid in industrialized societies for the process known as “individualization”, which has become a substantive marker of “reflexive modernity”. Indeed, some of the most prominent contemporary sociologists and philosophers, such as Gilles Lipovetsky, Elizabeth Beck, and Zygmunt Bauman, have advocated for this perspective. Strongly influenced by the sociological tradition of Norbert Elias, in his work *Liquid Modernity*, Bauman conceptualizes the phenomenon of “individualization” as a process that transforms human identity “from a given into a task, burdening actors with responsibility for this task and for the consequences (and side effects) of their actions” (Bauman 2003, 20).

The term “individualization” thus refers to the social process that is the consequence of the strong development of individualism that characterized late modernity in the second half of the twentieth century –qualified in sociological terms such as the above– mentioned “reflexive modernity”. Conversely, it is also the triumph of Libertarian logic during the first decades of the 21st century, which has led to the individual being regarded as the absolute owner and responsible for their own life. This process of reflection has rewarded the development of the capacity for self-determination above all else.

In the context of technological transformation of contemporary societies, sociologists like Ulrich Beck (2008) view “individualization” as a radical transformation of the personality structure of societies. In terms of Beck, this is because the isolated individuals are led to believe that we can be freed from the constraints of traditional societal structures, which enables us to have complete control over the development of our lives through the decisions we make in the processes of technological rationalization. However, the effect of this phenomenon of “individualization” is devastating: it results in the

¹ The present study is part of the MICINN “Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas” 2023-2025 (PID2022-136439OB-I00), supported by MCIN/AEI/10.13039/50110001103.

individual being detached from the community, weakening their trust in others until they are left unprotected in a virtual world in which it is increasingly difficult for people to develop in an autonomous way.

The extensive deployment of AI systems and digital media for surveillance and social control during the past few years has inevitably led us to face the “flip side of the coin”, namely the risk associated with AI tools (Beck and Gernsheim 2003). Do we need to question, then, the veracity of the anthropocentric philosophical and ontological paradigm, which posits that the human being is the only being in the world endowed with consciousness and that, consequently, human beings are the only ones who deserve ethical treatment?

In view of the profound impact that developments in biotechnology, neuroscience, and disruptive technologies such as AI could have on our understanding of the world in the coming decades, an increasing number of scholars are advocating for an approach to ethics in the context of techno-sciences that prioritizes the hermeneutical perspective by emphasizing the interdependence of knowledge through the integration of emotional insights. In this alignment, we endorse the proposal of the Spanish philosopher Fernando Savater (2011) who calls for abandoning the scientific reductionism of positivism and Heideggerian phenomenological factualism. Instead, Savater advocates for the hermeneutic-critical approach, which is well-known among other philosophers such as Habermas, Apel, or Adela Cortina (2007) in Spain. This approach aims to reinterpret the reality of the human person *vis-à-vis* their vulnerability in relation to technology (Cortina 2011).

In light of the growing acknowledgement during late postmodernity of the vulnerability of nature under the yoke of human technological intervention, the use of the term “responsibility” marks the transition from “inescapable reciprocity” between fellow human beings –meaning “love thy neighbor as thyself”– to a “responsibility towards nature” of a markedly teleological character. This transition is opposed to ethics based on conviction. Indeed, the Kantian categorical imperative –to act in such a way that the principle of one’s action becomes a universal law– could be adapted through new formulations of a collective imperative, through intergenerational action in the public sphere.

As observed by German philosopher Hans Jonas, it is accurate to conclude that AI will profoundly impact the world we live in. However, it will be ethical considerations that will ultimately shape the nature of this transformation. Indeed, the consequences derived from the use of disruptive technology exceed the traditional frame of ethics, forcing us

to question the “principle of responsibility” discussed by Jonas (1995) in his essay, *Ethics for Technological Civilization*. As this author notes, technological intervention has significantly altered the inescapable reciprocity aspect of the ethics of proximity by positioning nature at the service of humans, with profoundly detrimental and dehumanizing consequences. From Jonas’ perspective, the newfound capacity for human action on the natural world has fundamentally transformed the very nature of ethics. From the moment that human beings are able to destroy, they are held to a new standard with regard to future generations: namely, “the responsibility for what is to come”. This is the way in which the term “responsibility” is employed by Jonas, wherein responsibility is oriented towards the future and encompasses both the present and the past tenses.

The construction of a shared ethical framework for AI systems represents nevertheless a significant challenge, particularly in light of the postmodern phenomenon of ethical relativism. This challenges us to recognize that every community and individual may have different conceptions of what qualifies as morally acceptable. The collective approach previously outlined, in which Jonas’ responsibility for the future is considered, could provide a more explicit justification for the assumption of a “universal moral paradigm”. This is because it is based on shared human experiences and the enduring principles of philosophical traditions that have guided ethical grounding throughout history. Consequently, the question arises as to whether the construction of a shared ethical framework for AI is a utopian idea or rather an increasingly pressing need.

The establishment of a moral paradigm and a unified ethical framework for the advancement of AI is predicated on two fundamental assumptions. Firstly, as Fernando Savater notes, a universal ethical framework could serve as a safeguard against the shortcomings of the scientific reduction of positivism and factualism. Secondly, it encourages a hermeneutic approach that promotes interdependence through the transversality of the emotional approach in human knowledge, particularly in the context of technology (Cortina 2011).

In addition to the above, the proposal of a universal ethical frame of reference in AI facilitates a collective understanding of the boundaries that must not be transgressed by AI, which aligns with the fundamental purpose of technology. The term “technology” has its etymological roots in the Greek word *τέχνη* (*téchnē*), which signifies art, craft, or skill. From a broader perspective, technology is defined as a process or capacity to transform or combine existing elements in

order to create something new, thereby enabling the improvement and deepening of human existence. This approach to understanding technology is rooted in the Aristotelian concept that the value of goods, institutions, and social practices is contingent upon their intended purpose or end. The most accurate method for discerning the virtues –both ethical and dianoethical– appropriated to a process, craft, or skill, such as technology, when attempting to comprehend the *telos* of that process is the fundamental tenet of the Aristotelian Theory of Justice and the foundation of Virtue Ethics.

It is therefore imperative to engage in a comprehensive discussion regarding the ethical implications of technological advancements associated with the use of AI. This debate would entail a comprehensive examination of the potential harms and dangers that must be avoided, while also promoting those values of human interdependence that can establish an ecosystem of trust among citizens, stakeholders, and users in the face of a “potentially harmful use” of AI.

2. *EU Artificial Intelligence Act: A new roadmap on fundamental rights risk management in the face of AI*

For several years, the European Union has been engaging in a comprehensive strategy for responsible research and innovation in Techno-sciences, which is represented by the acronym RRI (Responsible Research and Innovation). The RRI program represents a novel approach to research governance, aiming to bridge the division between the scientific community and society. It encourages the socialization of techno-scientific environments, where civil society and technologists collaborate to align scientific research with societal values, needs, and expectations. More precisely, the RRI program encompasses six lines of action: (i) Citizen participation throughout the research process. (ii) Gender equality in work teams. (iii) Science education to improve educational processes and promote scientific vocations among the very young. (iv) Ethical awareness to foster scientific integrity, in order to prevent and avoid unacceptable research practices. (v) Free access to scientific information to improve open dialogue with society. And (vi) Governance agreements, with the aim of providing tools that foster shared responsibility among interest groups and institutions.

In this context, the EU strategy for an ethical and responsible program on research and innovation in techno-sciences has gained

particular prominence in recent years, particularly in view of the presentation of a harmonized system of rules within the EU in the field of Artificial intelligence, known as the *EU Artificial Intelligence Act* (EU AI Act). The EU AI Act represents a significant regulatory milestone in fostering a collaborative ecosystem of trust among citizens, stakeholders, and users in the context of AI, which has the potential to be employed in ways that may result harmful. It aims to establish a transnational AI regulatory framework, and it is the first cross-cutting legal regulation that is directly applicable in all EU Member States, eliminating the need for subsequent national transposition rules to be developed. Furthermore, the regulatory system established by the proposed EU AI Act is universal in scope, extending to all AI systems functioning as components of products or intended for placement within the European Union market, regardless of whether they are standalone AI systems or integrated components within larger products.

The initial aim of the EU AI Act was to control and manage risk by addressing deficiencies in existing legislation, with a view to establishing an effective risk-based approach to AI (Soriano 2021). Indeed, the proposal for a common European legal framework on AI incorporates a system based on risk management that establishes different information obligations for providers depending on the level of risk associated with the use of an AI system with respect to the guarantees of users' fundamental rights.

This distinction is reflected in the categorization of AI systems into three categories. The first level, AI systems of unacceptable risk (level A), is prohibited and applies to systems whose risk is so unacceptably high (see Title II). The second level, AI systems of high risk (level B), which is considered as high-risk systems, applies to systems that generate important risk or that could adversely affect the due guarantee and safeguarding of fundamental rights (see Title III). The third level, AI systems of limited risk (level C), applies to systems of limited risk that, though they are not considered high risk, have a series of transparency's requirements (see Title IV). There is also a fourth level for the remaining AI systems, which applies to all other permitted systems (see Title IX).

- i. The proposed EU AI Regulation establishes a first minimum obligation for those AI systems considered to be low risk or limited risk (Level C). Specifically, these AI systems require a minimum level of transparency's requirements that allow users to make informed decisions under their consent. Therefore, we

- would be dealing with a limited risk AI system when users are aware that the image, audio, or video content offered to them has been generated by an AI application or device. With regard to generative AI applications, such as ChatGPT, the EU AI Act proposes a special mention of the additional transparency requirements that must be met in order to be classified as “limited risk” applications. In particular, it imposes specific requirements for generative AI systems, including the following: (a) The content of AI system must be disclosed to the user as having been generated by an AI. (b) The AI system must publish periodic summaries of the copyrighted data used for training. (c) The AI system must prohibit the dissemination of illegal content.
- ii. AI systems that could adversely affect security or the due guarantee and safeguarding of fundamental rights are considered in the EU AI Act as high-risk systems (Level B). The European AI Act distinguishes between two categories of “high-risk AI systems” for the purpose of this distinction. (ii.1) The first category includes AI systems that are used in products subject to EU consumer product safety legislation –such as toys, aviation, automobiles, medical devices, or elevators–. (ii.2) Secondly, high-risk AI systems are defined as those that enable the following activities: (a) biometric identification and categorization of natural persons, (b) management and operation of critical infrastructure, (c) education and vocational training, (d) employment, management of workers and access to self-employment, (e) access to and enjoyment of essential private services and public services and benefits, (f) management of migration, asylum and border control, and (h) assistance in legal interpretation and law enforcement.
 - iii. Finally, the EU AI Act considers those systems that pose a direct threat to individuals and to the guarantee of their human rights as unacceptable risk AI systems (Level A) and expressly prohibits them. This prohibition extends to three essential modalities of AI systems: (iii.1) AI systems that employ cognitive manipulation of the behavior of vulnerable individuals or groups, such as children and adolescents; this prohibition encompasses the potential for AI devices to encourage dangerous behaviors in children or to induce suicidal behaviors in adolescents. (iii.2) AI systems that utilize algorithms to generate identity biases for the purpose of classifying individuals based on their socioeconomic status or personal characteristics, including race, gender, nationality, sexual orientation, religion, etc. (iii.3) Finally, AI systems that use

biometric identification, both in real time and remotely, which employ facial recognition.

In this regard, the arduous parliamentary discussions that took place among the Expert groups during the legislative process of reflection on EU AI Act have resulted in the extension of the list of AI systems to be considered prohibited to five new modalities: (a) Real-time remote biometric identification systems, when performed in public access spaces that would allow mass surveillance. (b) Delayed remote biometric identification systems, with the sole exception that the use of such systems are performed by state security forces and corps for the prosecution of serious crimes and by prior judicial authorization. (c) Predictive AI system that are able to anticipate the risk of committing criminal or administrative offenses. (d) Predictive AI systems that enable the inference of the emotions of a natural person in the domains of law enforcement and border management, in workplaces, and in educational institutions. (e) AI systems that employ subliminal techniques to materially distort the behavior of the same.

Despite the substantial support that the EU AI Act is expected to receive in the upcoming years, one of the most contentious issues that has emerged during the legislative process of the EU AI Act is the proposal to impose an appropriate level of prohibition on those AI systems that pose an “unacceptable risk” to the fundamental right of non-discrimination and to the safeguarding of ethical principles against the consolidation of negative stereotypes about religious or ethnic minorities. Bearing in mind the potential for discriminatory outcomes associated with AI systems, it is indeed crucial to give appropriate consideration to the stipulations laid out in Article 5 of the EU AI Act.

In accordance with Article 5.1a) of the EU AI Act, the utilization of any AI system that employs subliminal techniques or manipulative or deceitful methods for the purpose of influencing the behavior of an individual or group of individuals –and which is not discernible to the individual– is explicitly identified as an unacceptable practice. In such circumstances, the capacity of the individual to make an informed decision is significantly constrained, thereby increasing the likelihood of significant harm being inflicted on the individual or on another person. Notwithstanding the above, the prohibition of an AI system employing subliminal techniques shall not apply to AI systems intended for therapeutic purposes, on the condition that informed consent is obtained from the patients themselves or, when appropriate, from their legal guardians.

The ethical implications of this regulatory clause –as introduced in the final version of Article 5.1 a) of the EU AI Act– are significant. Accordingly, any attempt to influence our deep or unconscious mental processes through the use of subliminal techniques, or any manipulative or deceptive techniques employed in AI devices for the purpose of influencing our decisions as users or consumers about what to purchase, what to consume, what to appreciate, or what to despise, should be banned and declared null and void.

This stipulation pertains to all AI systems that prompt individuals to make decisions otherwise unmade by those individuals themselves. This disposition thus simultaneously targets two distinct forms of influence: (i) the manipulation of decision-making processes and (ii) the dissemination of disinformation with the potential to alter ethical, moral, or ideological convictions or identity, as well as religious beliefs. The second effect, which pertains to the role of misinformation in influencing opinions that may alter convictions or beliefs, is a particularly salient issue that warrants comprehensive investigation. This is particularly the case given that the EU AI Act has adopted a framework whereby it falls on the aggrieved party to prove damages. Firstly, the legislation mandates the presentation of evidence indicating that the decision in question would not have been reached by the user in the absence of the AI system. Furthermore, the EU AI Act stipulates that proof of a risk of significant harm must be provided, although it does not provide a definition for this term.

It follows that, should one accept the proposition that the human unconscious is worthy of legal protection, it is not adequate to prohibit only those deliberately deceptive or manipulative subliminal techniques used by AI systems with the intention of making a profit. This is not only because they have a considerable impact on the ability to make an informed decision, but also –as described in Article 5.1.a) of the EU AI Act– because it prompts us to consider whether the concept of “own and voluntary act” is called into question. In the context of our study, it becomes evident that a legal framework is required to protect individuals from exploitation by those seeking to influence their actions below the level of conscious awareness. It thus follows that a framework of protection that is conducive to the human condition within AI systems is of great necessity.

An additional legal issue that presents analogous challenges pertains to the prohibition established in Article 5.1(b) of the EU AI Act. This article prohibits those AI systems that exploit the vulnerabilities of individuals based on their age, disability, or specific social or economic circumstances with the objective of materially distorting an individual’s

behavior in a way that may cause them significant harm. The second provision of Article 5 has a considerable scope of application, encompassing AI systems that interact directly with users, such as chatbots or recommendation-based AI systems.

Moreover, identifying the areas of vulnerability that may bring an AI system within the scope of the prohibition set forth in Article 5.1(b) of the EU AI Act is also a challenging endeavor. This is because there is no definition of the term “vulnerability” in the Act itself or in relation to each of the characteristics listed in Provision 5.1 (b). Therefore, a broad interpretation would result in the prohibition of manipulative systems that exploit the vulnerabilities of specific groups with the intention of modifying their behavior and causing them harm or damage.

It is clear that this regulatory structure is inadequate because of the conceptual ambiguity surrounding the concept of vulnerability, which impedes the effective implementation of Provision 5.1(b). Furthermore, the legal approach places the burden of proof on vulnerable users, which is a further shortcoming. Firstly, it is essential to establish whether the primary objective of the AI system in question is to exploit the vulnerabilities of a specific demographic. Secondly, it is necessary to demonstrate that the AI system in question actually distorts the behavior of the aforementioned vulnerable individuals, rather than simply appearing to do so. In practical terms, an individual seeking to demonstrate that a particular AI system has deliberately exploited a specific vulnerability must address the considerable challenge of assembling a compelling body of evidence to prove the system’s malicious intent.

In particular, the necessity of implementing specialized safeguarding measures for cognitive freedom in the context of generative AI systems becomes a pressing need. It is imperative that a commitment be made to develop a consensual international legislative framework that subordinates the design, production, and development of AI to the dignity of the human condition. Secondly, a comprehensive set of protections is required to safeguard the human condition. This necessitates the criminalization of any deliberately manipulative or deceptive techniques employed by the implementation of AI applications or devices with the objective of influencing the ethical and critical consciousness of individuals, thereby limiting their ability to make autonomous decisions and actions. Thirdly, it is imperative to acknowledge the necessity of recognizing the existence of a fundamental cognitive freedom, which is indispensable for the safeguarding of the unconscious mind. The unconscious mind is the

inalienable foundation of human individuality and, as such, requires protection. This would entail establishing the source of validity for the requirement of “free consent” in any legal act. The governance of AI systems is likely to become the exclusive domain of political and economic elites in the near future. This may well result in increased inequalities and the emergence of new mechanisms of social exclusion. It is therefore of the utmost importance that regulatory measures be put in place to prosecute and punish the use of AI as a device for social control and manipulation.

Furthermore, with the growing prevalence of automated and generative AI systems comes a concomitant increase in instances of algorithmic discrimination. In these cases, the use of AI perpetuates the vulnerabilities faced by specific disadvantaged groups and minorities. Indeed, the advancement and implementation of IA systems founded on the use of algorithms present a multitude of possibilities for the incorporation of biases that are detrimental to members of disadvantaged groups. One of the means through which algorithms may perpetuate the traditional discriminatory structures faced by minorities is through the selection and weighing of variables employed by IA systems for the measurement and prediction of the object under consideration.

The order of priority assigned to specific variables in the measurement of the phenomena predicted by algorithms can affect the outcome of these programs. To illustrate, if a bank’s customer credit rating system prioritizes income level over savings capacity as an indicator of an individual’s creditworthiness, this decision will result in a greater disadvantage for women, millennials, immigrants, and other vulnerable groups.

Humans may be particularly susceptible to the influence of AI due to the potential psychological harm caused by technology that seeks to exert control over human will. In recognition of the potential for exploitation of human vulnerability in contexts of technological disruption, the European AI Act acknowledges the necessity for accountability and protection of individuals in the context of AI-driven change. In light of these considerations, the EU AI Act prohibits the exploitation of vulnerabilities of groups of people on the basis of their age, disability, or social or economic situation. This is done in a manner that distorts their behavior and is likely to cause harm to them or to others. Furthermore, in the context of safeguarding fundamental rights to equality and non-discrimination, the EU AI Act strives to strike an appropriate balance between the advancement of individual autonomy and economic efficiency (Soriano 2023).

3. Preventive risk control: A contemporary form of discrimination

It has to be noted that artificial intelligence systems are not immune to the biases and prejudices that exist in society. Such biases tend to pervade the algorithms themselves, thereby facilitating the propagation of discriminatory outcomes. Consequently, an unbalanced or inappropriate selection of data during the training of an AI system may result in the algorithm making unfair decisions that lead to the stigmatization of certain groups, minorities, and/or individuals. Such bias may be generated by a number of factors, including preconceived beliefs, predilections, or unconscious prejudices that have been acquired by individuals throughout their lives based on the sociocultural stereotypes they may have acquired at different stages of life.

The *Rome Convention for the Protection of Human Rights and Fundamental Freedoms* of November 4, 1950, contains in Article 14 an “anti-discrimination clause » which encompasses both a general equality clause and a clause prohibiting discrimination on certain specific grounds, including gender, race, and ethnic origin. Furthermore, the Council of Europe has endeavored to remove the limitations imposed by the current Article 14 of the Convention through the approval of Protocol 127, which recognizes a broad prohibition of discrimination. In particular, the first article of the Explanatory Report to Protocol 127 states: “The exercise of any right recognized by law shall be secured without any discrimination based, in particular, on sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, wealth, birth or other status.”

The European Court of Human Rights (HUDOC) defines the term discrimination in the case *Willis v. United Kingdom* (September 11th, 2002) as “treating differently, without objective and reasonable justification, persons in substantially similar situations”. In a related case, *Thlimmenos v. Greece* (April 6th, 2000) the Court broadened the scope of this clause to encompass discrimination by indifference, which refers to the existence of discrimination when states do not treat differently, without objective and reasonable justification, persons whose situations are substantially different (McCrudden 2008, 712-724). In accordance with the HUDOC doctrine, there would be discrimination when individuals are treated identically under the law, yet their circumstances differ –that is, discrimination by differentiation– and when individuals are treated differently despite their comparable circumstances –that is, “discrimination by indifferenciation”. Nevertheless, it seems unlikely that

this doctrine of “discrimination by indifferenciation” will have a significant future impact. However, examples of the latter can be found in the use of a type of algorithms used in AI applications for facial recognition, particularly *ante facto* predictive algorithms, which show a significant discriminatory bias when identifying people of different racial and ethnic origins (Berk et al. 2018, 1-24).

In light of the European legal system’s foundation on liberal premises regarding the advancement of individual autonomy, the traditional European legal framework in the domain of equality and non-discrimination has historically integrated two distinct categories of legal instruments to safeguard individuals against discriminatory practices. (i) The initial type is a preventive or *pre-facto* instrument against discrimination. This is also known as an anti-classification legal instrument. (ii) The second type of legal instrument is a reactive or *post-facto* instrument. This is also known as an anti-subordination instrument (Ganti and Benito 2021).

The anti-classification legal instruments –also known such as *ante facto* prevention instruments– refer to those rules that prohibit the consideration of particularly suspect categories in decision-making processes. Consequently, these are reactive legal instruments in the face of situations of discrimination. In addition, anti-discrimination prohibitions function to some extent as preventive mechanisms, articulating mandates that seek to avoid discriminatory decision-making such as article 9 of the General EU Data Protection Act, which prohibits the processing of special categories of personal data, including racial origin, religious convictions, or political opinions. In contrast, the second type of anti-subordination legal instruments –also known such as *post-facto* repression instruments– seek to reverse those social structures that place persons belonging to certain groups or minorities in situations of disadvantage or discrimination. Consequently, these *post-facto* mechanisms aim to identify and redress all kinds of violations of the general non-discrimination principle.

The use of AI-based predictive algorithms in surveillance systems presents a significant challenge to the exercise of social control in the context of digital environments. Indeed, the entire community is placed under the dependence of a single criterion: *risk control*. The function of *ante facto* predictive algorithms is to determine the possible degree of risk posed by a person throughout the different stages of the criminal process, specifically in regard to the possibility of recidivism. From a purely normative perspective, the utilization of predictive risk algorithms presents significant challenges for legislators tasked with the legal regulation of long-term stability, the effective repression and

combating of violations of fundamental rights, and the pursuit of justice (Añón 2022, 17-49). Furthermore, it presents a challenge for legal professionals, as predictive AI systems are designed to anticipate human behavior, which could potentially result in the formation of discriminatory biases based on such factors as gender, nationality, ethnic origin, race, or religion.

A case closely aligned with the current topic of discussion is that of the Risk Indication System, which is also known by its acronym, SyRI System. This example will demonstrate the discriminatory implications of *ante facto* risk models. The SyRI System was used by the Dutch government to prevent and combat social security benefit fraud. This system allowed the Dutch public administration to use risk reports for claimants of child benefits in preventing the illegal obtaining of government funds in the field of social security. The SyRI System was established on the basis of the normative framework provided by national law, the so-called *Law on the Structure of Work and Income Enforcement Organization*, which contains in article 65.2 an extensive list of categories of information that may be processed in the SyRI system: in particular, gender, employment history, taxes, property ownership, education, health insurance, government permits, level of debt, track of public benefits received, and administrative sanctions such as traffic fines. To calculate potential evasion and fraud irregularities, the SyRI System algorithms linked all the applicants' personal data stored by government agencies and matched them with a "risk profile" generated from the information of other citizens with criminal records. Once any similarities and/or discrepancies were established, the system produced risk reports on a list of names as "potential fraudsters" that could be retained by the authorities for up to two years. Additionally, The SyRI System was substantiated in neighborhood projects in which government agencies identified those municipal districts most adequate to implement this risk assessment system: in practice, the poorest neighborhoods and municipal districts characterized by high rates of immigrant population. As a result, the Dutch administrative authorities wrongly accused hundreds of families receiving benefits of fraud simply because of their Moroccan or Arab origin.

This SyRI case prompted a landmark judicial precedent in Europe, which resulted in the first court decision to examine an algorithmic risk assessment system. The Netherlands Committee of Jurists for Human Rights v. State of the Netherlands is a judgment issued on March 6, 2020, in which the court concluded that the SyRI system had not only affected the human right to privacy, but also violated the transparency

requirement of Article 8 of the European Convention on Human Rights. Moreover, the court examined the legitimacy of the government's use of citizens' risk reports to determine the allocation of social benefits. It concluded that the SyRI system was "neither transparent nor verifiable" not only because such a system could be used to create data profiles of individuals for other purposes, which are prohibited by law, but also because the risk models used by the Dutch government were never published. The interested parties were not notified in advance of the above when their data were entered into the SyRI system for the preparation of their risk profile before the public administration. Indeed, with regard to the balancing test, the court determined that a risk report has a non-negligible legal effect on the right to privacy of the individual subjected to algorithmic scrutiny, because such a report cannot preclude the use of sensitive information in subsequent procedures and communications between citizens and the public administration. Based on this reasoning, the court dismissed the "declared interest of the Dutch government".

Nevertheless, it should be acknowledged that when algorithms implement discriminatory practices based on so-called "suspect categories" or "ante facto prevention categories", they often employ a non-maleficence approach, whereby ostensibly impartial measurement criteria are, in reality, utilized in a manner that ultimately results in the disadvantage of individuals belonging to ethnic and racial minorities when compared to their non-minority counterparts. We shall now proceed to provide further clarification on this matter.

4. The principle of *Non-maleficence*: Prevention of harm and preservation of human dignity in the face of the risk of AI

In 1979, two distinguished American philosophers, Tom Beauchamp and James Childress, published a seminal work entitled *Principles of Biomedical Ethics*, which laid the groundwork for contemporary discourse within the field of ethics applied to medical sciences. In this publication, the aforementioned philosophers put forth four ethical principles as follows: (i) respect for autonomy, (ii) the principle of non-maleficence, (iii) the principle of beneficence, and (iv) justice. The authors presented these four principles, which have long been observed in human societies and have governed ethical behavior, as applicable to any culture or society (Beauchamp and Childress 1994).

The principle of non-maleficence is rooted in the classical medical maxim *primum non nocere*, which can be translated to "first do no

harm". It refers to the ethical obligation of avoiding any intentional infliction of harm. Indeed, the principle of non-maleficence can be defined as the obligation not to cause harm or to prevent harm from occurring. It encompasses the prohibition against killing, inflicting pain or suffering, and causing disability. Such a breach constitutes a public wrongdoing and is therefore subject to legal consequences.

Moreover, there is a clear distinction between the principle of not inflicting harm upon others, which encompasses behaviors such as theft and murder and the obligation of beneficence, which aims to safeguard personal interests or advance the collective good. When applied to AI systems, the principle of non-maleficence would ensure that such systems prioritize the safety of individuals or prospective users, as well as the preservation of human dignity. Consequently, this principle would serve to reduce risk and enhance transparency and explainability. More precisely, there are numerous instances in which AI devices have already incorporated the principle of non-maleficence with the objective of enhancing user safety. A case that exemplifies this concept can be observed in the automotive industry, particularly in the integration of AI systems into autonomous vehicles with the objective of reducing traffic accidents and enhancing road safety. The deployment of autonomous vehicles has the potential to result in a significant reduction in accidents caused by human error, which include driver inattention, visual fatigue, and lack of reflexes. Given the goal of AI devices applied to autonomous vehicles of reducing potential harm while simultaneously maximizing the safety of individuals on the road, such an approach would thus represent an instance of the implementation of the principle of non-maleficence.

In contemplating the possible applications of the non-maleficence principle to AI systems, it is *prima facie* necessary to examine the role of virtues such as kindness, empathy and compassion in the machine learning of AI systems, in natural language processing for the design of AI applications that are able to perceive, understand and respond to human emotions. Indeed, in the process of machine learning, AI systems are capable of processing vast quantities of data in order to make predictions about human behaviors that have already occurred. It is not thus simply a matter of creating intelligent machines that can replace humans in their reasoning and cognitive capacities. Instead, it is about fostering virtuous AI that reflects the best of human moral aspirations.

Consequently, if AI systems are trained to collect data that exemplifies the essential virtues to the human condition –such as kindness, empathy, solidarity, courage, prudence, and compassion– they can thus be enhanced with the ability to recognize and respond to

situations in ways that promote the common good in human communities. Machine learning offers an endless array of possibilities for instructing AI in the moral obligation to prevent or alleviate harm –therefore, to do good– and in the duty to help others over and above private interests. In other words, to act for the greatest possible benefit, seeking the greatest possible general welfare.

5. **Mitigating the discriminatory impact of biases in AI algorithms: Seeking the beneficence principle**

The term “beneficence” is generally accepted as the act of performing benevolent acts or actions that are perceived to be beneficial to others. Beyond the necessity to abstain from causing harm to others, the principle of beneficence obligates individuals to demonstrate concern for, and actively promote the well-being of, those around them. Indeed, the term “beneficence” is generally understood to encompass a broad range of behaviors, including acts of mercy, kindness, charity, altruism, love, and humanity.

Moreover, as defined by Beauchamp (2003, 12), beneficence encourages individuals and institutions to feel an ethical obligation to contribute actively to the welfare of the community by promoting civic virtues such as altruism, solidarity, compassion, and social responsibility in human actions. This principle implies a beneficial action that prevents or counteracts evil or harm, and additionally confirms the absence of acts that could cause harm.

The principle of beneficence furthermore represents a fundamental tenet within ethical theories, including the moral doctrine of utilitarianism. This is evident in the formulation of the utility principle, which states that actions should be taken to promote the welfare and act in a way that maximizes the happiness of the greatest possible number of people. This approach to the shortcomings of utilitarian reasoning is particularly evident in the context of AI, as it necessitates the focalization of efforts on mechanisms of beneficence that allow for the mitigation of unnecessary harms associated with AI systems, especially those that could significantly compromise collective welfare. It is therefore imperative that any AI system developers address the issue of mitigating the impact of biases in AI algorithms in order to comply with the principle of beneficence. Let us elaborate further on the concept of beneficence, which enables the mitigation of unwarranted damage caused by AI systems, particularly those with the potential to significantly harm collective welfare.

There are no *ex ante* regulatory control mechanisms to ensure that AI systems are not discriminatory. Indeed, one of the primary limitations of the European legal framework is that mechanisms against discrimination typically operate *ex post*, that is, after the discriminatory action has already occurred. From a purely normative perspective, the implementation of *ante facto* predictive AI systems presents significant challenges for the legislature in its efforts to effectively repress situations that violate fundamental rights (Gerard and Xenidis 2021).

Algorithmic discrimination may also originate from errors or biases present in the databases utilized in the development of automated decision-making systems, as AI systems using predictive algorithms are designed with data related to the phenomenon they seek to predict. Once the system has been trained, its performance will be evaluated with data used to detect its level of accuracy. For instance, a database of arrests and convictions may contain primarily data on individuals from ethnic and racial minorities as a consequence of the pervasive discrimination they have historically faced in their interactions with law enforcement and the justice system. In this instance, the algorithm would be encouraged to learn that certain persons belonging to certain ethnic or racial minorities are more likely to engage in criminal activities. In other cases, the use of algorithms may serve to perpetuate stereotypes that underpin social structures of discrimination (Makonnen 2007). For example, the results yielded by entering specific combinations of words into Internet search engines, such as Google, have been found to reproduce gender roles or at least to contribute to the consolidation of negative stereotypes about religious or ethnic minorities (García-Berrio 2023). If the data used to train an AI system is of poor quality, the result may be that the algorithms induce us to undertake decisions that result in stigmatization of certain groups or minorities. This is due to the so-called “biases”, which are understood to be preconceived beliefs, predilections, or unconscious prejudices that have been acquired throughout one’s lifetime based on the sociocultural stereotypes with which one has been educated (Castellanos 2023).

Furthermore, predictive AI systems pose a challenge for legal professionals, as they are designed to identify patterns in human behavior. As we have pointed out, this may lead to the creation of discriminatory biases based on factors such as gender, nationality, ethnic origin, race, or religion. In light of the historical structures of discrimination that have placed certain ethnic groups and religious minorities in positions of disadvantage or subordination, it is evident

that when an AI system employs *ante facto* predictive algorithms, the validation and test data utilized to train the AI system would probably reflect historical structures of discrimination based on race, gender, religion, etc. As a result, the system may assume that the biases it contains are accurate or valid. It has, in fact, been demonstrated that the current bias produced by the use of AI systems is due to the imbalanced representation of ethnical traits that developers of AI systems employ in the training data (Žliobaitė and Custers 2016). This representation will tend to include, for instance, a greater number of male and light-skinned faces. Conversely, if AI systems are employed as a predictive tool to generate profiles of potential perpetrators of a homicide, the validation data set will contain information related to homicides that have already been solved. Consequently, male and dark-skinned faces would predominate. For instance, a series of predictive risk algorithms have been implemented in recent years that can be applied to persons who respond to criminal stereotypes associated with different racial groups or ethnic origins. These stereotypes may increase the perception of guilt. This was exemplified by the long-standing use of the COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) model in the United States to assess the likelihood of recidivism among criminal justice system participants. The COMPAS system revealed overt algorithmic discrimination against African-American males, who represent the majority of U.S. prison population with the longest prison sentences –including life imprisonment– of any other racial and gender combination. The COMPAS system was found to exhibit a pronounced racial bias as far as African-American males being more likely to be misclassified as higher risk –a profile that was reflected in their COMPAS scores–. This flagrant racial bias in the COMPAS system has received considerable public attention, prompting concerns about the potential discriminatory impact of AI algorithms when used in jurisdictional decisions in the criminal justice system.

In light of the discriminatory impact of biases in predictive AI algorithms, article 44 of EU AI Act outlines a number of mandatory requirements to be met by the training, validation, and test data of high-risk AI systems, as well as by the individuals or organizations responsible for collecting such data and processing it. Article 44 of the proposed EU AI Act states that “high data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are used”. The objective is to ensure that the high-risk AI system performs as intended and safely, and that it does not become the source of discrimination prohibited by EU AI Act

(Hacker 2018). In addition to prior regulation, the second paragraph of Article 10 states that “training, validation, and testing data sets shall be subject to appropriate data governance and management practices”. The fourth paragraph of Article 10 continues to elaborate the above: “Training, validation, and testing data sets must be considered in accordance with the intended purpose, taking into account the specific geographical, behavioral, and functional characteristics of the environment in which the high-risk AI system is intended to be used”.

Despite the multitude of challenges to the principle of beneficence in the context of potential discrimination due to algorithmic bias, there is still a glimmer of hope. In light of the above, it is encouraging to report that a new guarantee of “human supervision” has been introduced in the EU AI Act. This new measure requires those responsible for the management of AI systems to be aware of the risks associated with bias, automation, or confirmation of the potential issues inherent in this type of digital application. In this regard, the European Parliament calls upon managers of AI systems to comply with their legal obligation to provide specifications for the input data or any other relevant information regarding the data sets used in AI systems, taking into account the intended purpose and the reasonably foreseeable misuse of the system.

Indeed, the EU AI Act introduces a new ethical duty favoring a recognition of the pivotal role played by intersubjectivity and the human condition within AI systems. This allows us to highlight the main disadvantage of AI: artificial intelligence and its algorithms lack the capacity to feel and possess no moral conscience. They are capable of understanding, but not of comprehending.

Conclusions

Increasing affective learning in automated AI processes serves to augment the capacity of AI systems to discern, comprehend, and respond to the nuances of human emotion. Nevertheless, as has been discussed in this paper, it is imperative to exercise caution in the acceptance of such advances. While AI systems may demonstrate empathic, kind, and compassionate behavior, they certainly lack the emotional connection that derives from human experience. This can be articulated in Kantian terms as the condition of humanity.

Any effort to establish an ethical framework for AI has the potential to imbue technology with a humanizing quality through the

promotion of virtues intrinsic to the human condition. This notable dedication to incorporating human factors into AI processes highlights the pressing concerns associated with the integration of AI in our daily lives. Consequently, integrating the three ethical virtues of kindness, empathy, and compassion into the configuration of AI systems paves the way for the creation of AI with a profound sense of humanity.

As our research illustrates, the ethical quality of AI systems can be enhanced through the implementation of fairness, which in turn facilitates the acknowledgement of the principle of non-maleficence. This is merely a method of circumventing the potential detriment that may result from the implementation of algorithmic biases. In addition, the promotion of empathy requires AI system developers to make use of the beneficence principle to offset hidden biases in *de facto* risk-avoidance algorithms, as well as discriminatory effects based on ethnic and racial identities that inevitably result in the segregation of minorities. In conclusion, the implementation of compassion into AI represents the pivotal impetus behind the mounting ethical pressure vis-à-vis the accountability of AI programmers and developers in the context of biased algorithmic sequences and the malevolent consequences of some of the latest generative AI utilities.

Consequently, any attempt to build an ethical framework for AI should acknowledge and accept the moral responsibility of human beings. The integration of kindness, empathy, and compassion into the design of AI systems would allow for the decisive prioritization of users' welfare, the promotion of fairness, transparency, and accountability, and the assurance that AI technologies serve the interests of citizens rather than those of technology corporations or *de facto* powers. Such a perspective should inform the development of AI algorithms that prioritize empathy, respect, and human dignity over the construction of discriminatory biases. Indeed, as we imbue machines with intelligence and decision-making capacity, the virtues we can instill in them become the very cornerstone of the ethical development of technology, especially in regard to addressing the potential systemic injustices that could result from a variety of biases in data and discriminatory algorithms.

In this study, we have selected a number of examples –including the SyRI system– with the intention of illustrating the primary challenge that the EU AI Act presents in terms of the automation of algorithms employed by predictive *ante facto* risk control systems. Indeed, the use of predictive AI systems by governments to generate “risk reports” for their citizens calls into question one of the epistemological foundations of the legal definition of the rule of law, namely the autonomy and self-determination of individuals.

Those who adhere to contemporary interpretations of self-determination employ the Kantian ideal of moral autonomy to challenge the perspective of those who vehemently criticize the libertarian conception of personal autonomy as individualistic (or even selfish). In essence, Libertarians prioritize personal autonomy over subjectivities and preferences, a stance that is detrimental to the common good. For this reason, postmodern liberal thinkers such as Robert Young (1980, 573–576) and Joseph Raz (1986, 373) have proposed the notion of socialized autonomy, which effectively synthesizes the classical Kantian ideal of autonomy of the will with challenges such as those posed by the new applications of algorithms in AI systems. If our autonomy and ability to act freely are compromised through the use of predictive algorithms like the SyRI System, we no longer act according to a maxim that we have chosen for ourselves, but in compliance with a maxim that the community must establish for the common good.

Notwithstanding these limitations, it is crucial to acknowledge that one of the primary allures of AI applications is their capacity to present themselves as a means of overcoming human subjectivity, or even of eradicating stereotypes and social prejudices. The appeal to the certainty and neutrality of algorithms is an effective method for gaining acceptance and trust. Nevertheless, as illustrated by the SyRI System, there is a potential risk for algorithmic systems to be exploited by public agencies with the intention of establishing a repressive system that would be detrimental to public freedoms and fundamental rights, including the freedom of belief and the freedom of thought. Furthermore, the constitutional rights of citizens may be violated by the use of certain AI risk assessment systems, and human dignity may be infringed upon, particularly in the case of ethnic minorities and immigrant populations (Zuboff 2020). As previously observed, any individual subject to algorithmic scrutiny could potentially be prosecuted on the grounds of behavioral predictions generated by algorithms that may be perceived as risky. Rather than being prosecuted for the acts committed, individuals would be prosecuted *ex ante* based on identity biases associated with algorithms that take into account a range of factors, including an individual's level of income and indebtedness, their interactions on social networks, their place of residence, their religion, or their ethnic origin.

At the present time, the commendable attributes of AI systems are upheld on the basis of their reliability and predictability. However, it is crucial to consider the potential for predictive AI systems to be exploited for malevolent purposes, which could result in the consolidation of

authoritarian forms of governance and the undermining of democratic and citizen engagement. In this context, the construction of algorithmic patterns enabling machines to anticipate human behavior based on an ex ante predictability undermines the very concept of human autonomy. Accordingly, when adopting an ontological stance that recognizes the pivotal role of moral intersubjectivity and human autonomy, it becomes necessary to evaluate the primary limitation of predictive AI. This is because algorithms are unable to perceive human emotions and possess no moral conscience.

Current developments in the regulation of AI, as exemplified by the EU AI Act, posit human beings as the sole entity endowed with consciousness and the capacity to act autonomously. If we accept the proposition that autonomy and self-determination, in addition to the human conscience, are to be protected as a legal asset, we should consequently align ourselves with the European legislators. In that case, any recourse to predictive risk techniques employed by AI systems that have a significant impact on the aforementioned fundamental rights –including cognitive freedom, freedom of thought, belief, and religion– must be declared null and void.

References

- Añón, María José. 2022. «Desigualdades algorítmicas: Conductas de alto riesgo para los derechos humanos.» *Derechos y Libertades* 47 (1):17-49.
- Bauman, Zygmunt. 2003. *Modernidad Líquida*. Buenos aires: Fondo de Cultura Económica.
- Beauchamp, Tom. 2003. «The nature of applied bioethics.» In *A Companion to applied ethics*, edited by Roger Frey & Christopher H. Wellman, 1-16. Malden: Blackwell Publishing.
- Beauchamp, Tom and James Childress. 1994. *Principles of biomedical ethics*. New York: Oxford University Press.
- Beck, Ulrich. 2008. *La sociedad del riesgo mundial: En busca de la seguridad perdida*. Barcelona: Paidós.
- Beck, Ulrich and Elisabeth Gernsheim. 2003. *La individualización: El individualismo institucionalizado y sus consecuencias sociales y políticas*. Barcelona: Paidós.
- Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth. 2018. «Fairness in criminal justice risk assessments: The state of the art.» *Sociological Methods and Research* 50 (1): 1-24.
- Castellanos, Jorge. dir. 2023. *Inteligencia artificial y democracia: Garantías, límites constitucionales y perspectiva ética ante la transformación digital*. Barcelona: Atelier.

- Cortina, Adela. 2007. *Ética de la razón cordial: Educar en la ciudadanía en el siglo XXI*. Madrid: Nobel.
- Cortina, Adela. 2011. *Neuroética y neopolítica: Sugerencias para la educación moral*. Madrid: Tecnos.
- Ganty, Sarah and Juan Carlos Benito. 2021. *Expanding the list of protected grounds within anti-discrimination law in the EU*. Brussels: Equinet.
- García-Berrio, M^a Teresa. 2023. «La sociedad digital como cultura del riesgo: Desafíos éticos e implicaciones legales del uso de sistemas de Inteligencia artificial para la evaluación de riesgos y la vigilancia preventiva.». In *Inteligencia artificial y Democracia: Garantías, límites constitucionales y perspectiva ética ante la transformación digital*, edited by Jorge Castellanos, 39-65. Barcelona: Atelier.
- Gerards, Janneke and Raphaela Xenidis. 2021. *Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. European network of legal experts in gender equality and non-discrimination*. Luxembourg: European Union.
- Hackers, Philipp. 2018. «Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law.» *Common Market Law Review* 55 (4): 1143-1186.
- Jonas, Hans. 1995. *El principio de responsabilidad: Ensayo de una ética para la civilización tecnológica*, Barcelona: Herder.
- McCrudden, Christopher. 2008. "Human dignity and judicial interpretation of human rights". *European Journal of International Law* 19 (4): 655-724. doi.org/10.1093/ejil/chn043
- Makonnen, Timo. 2007. *Measuring Discrimination: Data collection and EU Equality Law: Thematic Report of the Group of Independent Experts*. Brussels: European Commission. Access December 9, 2024: <https://www.tandis.odihr.pl/bitstream/20.500.12389/19825/1/03245.pdf>
- Raz, Joseph. 1986. *The morality of freedom*. Oxford: Clarendon Press.
- Savater, Fernando. 2011. *Ética a Amador: Una invitación a vivir sin odio ni miedo*. Barcelona: Ariel.
- Soriano, Alba. 2021. «La propuesta de Reglamento de Inteligencia Artificial de la Unión Europea y los sistemas de alto riesgo.» *Revista General de Derecho de los Sectores Regulados* 8 (1): 50-63.
- Soriano, Alba. 2023. «Creando sistemas de Inteligencia Artificial no discriminatorios: Buscando el equilibrio entre la granularidad del código y la generalidad de las normas jurídicas". *IDP Revista De Internet, Derecho y Política* 38: 1-12. doi:10.7238/idp.v0i38.403794.
- Young, Robert. 1980. «Autonomy and Socialization». *Mind* 89 (356): 565-576.
- Žliobaitė, Indre and Bart Custers. 2016. «Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models.» *Artificial Intelligence & Law* 24 (2): 183-201. doi: <https://doi.org/10.1007/s10506-016-9182-5>.
- Zoboff, Shosana. 2020. *La era del capitalismo de la vigilancia: La lucha de un futuro humano frente a las nuevas fronteras del poder*. Barcelona: Paidós.