IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

https://idp.uoc.edu

ARTICLE

Creating non-discriminatory Artificial Intelligence systems: balancing the tensions between code granularity and the general nature of legal rules

Alba Soriano Arnanz
Universitat de València

Date of submission: September 2022
Accepted in: March 2023
Published in: October 2023

Abstract

Over the past decade, concern has grown regarding the risks generated by the use of artificial intelligence systems. One of the main problems associated with the use of these systems is the harm they have been proven to cause to the fundamental right to equality and non-discrimination. In this context, it is vital that we examine existing and proposed regulatory instruments that aim to address this particular issue, especially taking into consideration the difficulties of applying the abstract nature that typically characterises legal instruments and, in particular, the equality and non-discrimination legal framework, to the specific instructions that are needed when coding an artificial intelligence instrument that aims to be non-discriminatory. This paper focuses on examining how article 10 of the new EU Artificial Intelligence Act proposal may be the starting point for a new form of regulation that adapts to the needs of algorithmic systems.

Keywords

algorithms; equality; discrimination; biases; artificial intelligence

Creando sistemas de inteligencia artificial no discriminatorios: buscando el equilibrio entre la granularidad del código y la generalidad de las normas jurídicas

Resumen

En la última década ha crecido la preocupación por los riesgos que genera el uso de sistemas de inteligencia artificial. Uno de los principales problemas asociados al uso de estos sistemas son los riesgos que su uso genera para el derecho fundamental a la igualdad y a la no discriminación. En este contexto, debemos examinar los instrumentos normativos existentes y propuestos que pretenden abordar dichos riesgos, prestando especial atención a las dificultades de aplicar el carácter abstracto que suele caracterizar a las normas jurídicas y, en particular, al marco jurídico de la igualdad y la no discriminación, a las instrucciones específicas que se necesitan a la hora de programar un sistema de inteligencia artificial no discriminatorio. Este trabajo se centra en examinar cómo el artículo 10 de la nueva propuesta de Reglamento de Inteligencia Artificial puede ser un punto de partida útil en el camino hacia una nueva forma de regular que se adapte a las necesidades de los sistemas algorítmicos.

Palabras clave

algoritmos; igualdad; discriminación; sesgos; inteligencia artificial

Introduction

Discrimination caused or mediated by the use of automated systems has been recognised by the scholarship and institutions as one of the main risks arising from the growing use of algorithms in many areas of economic and social life (Gerards & Xenidis, 2021). In this context, over the past few years, a field of research specifically focused on investigating the development of systems respectful of the equality principle has emerged within the growing body of work related to algorithmic discrimination (Bent, 2020; Berk *et al*., 2018; Chouldechova, 2016; Corbett & Goel, 2018; Friedler *et al*., 2018; Pleis *et al*., 2017). Published works in this area propose mechanisms to incorporate "equality by design" into Artificial Intelligence (hereinafter, AI) systems (Renan-Barzilay & Ben-David, 2017, p. 430), while attempting to establish a formula that defines what constitutes a non-discriminatory system.

However, this line of research has not been able, to date, to give a conclusive answer as to the parameters that should be introduced in an automated system to ensure respect for the rights to equality and non-discrimination.[1] One of the main reasons why it is extremely difficult to establish fixed criteria with which all automated systems must comply in order to be considered non-discriminatory is the abstraction that characterises legal norms. In this sense, given that the interpretation and application of the normative framework for the protection of the rights to equality and non-discrimination is carried out on a case-by-case basis, there is a significant variation in the applicable criteria depending on the context, the type of decision made and the people affected, among other elements (Wachter *et al*., 2021).

Notwithstanding the existence of some rules applicable to the use of automated systems, such as article 22 of the General Data Protection Regulation (hereinafter, GDPR), which generally prohibits decisions solely based on the automated processing of data, we do not yet have a regulatory corpus designed to address, in a comprehensive manner, the different problems and risks generated by the growing use of automated systems. For this reason, most scholars have focused on analysing how regulatory instruments in the fields of transparency, equality and data protection, amongst others, can be applied to the use of artificial intelligence. Much of this work has focused on the inadequacies of existing rules to address some of the challenges posed by the use of AI systems (Cerrillo i Martínez, 2019; Huergo Lora, 2020; Valero Torrijos, 2020).

The aim of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (hereinafter, proposal for an EU AI Act) is to address the inadequacies of the existing rules as well as to establish an effective control system for AI systems (Soriano Arnanz, 2021a). This regulatory proposal establishes a risk-based approach and establishes four levels of risk:

- Systems whose risk is so unacceptably high that they are prohibited (Title II);
- Systems that generate high risk (Title III);
- Systems to which, though they are not considered high risk, a series of transparency requirements apply (Title IV);
- The remaining systems (Title IX).

Most of the proposal for an EU Artificial Intelligence Act focuses on the regulation of high-risk systems. Within the requirements to be met by these systems, this paper focuses on those contained in article 10, which refer to the data used in the "training, validation and testing" of high-risk systems. Thus, the aim of this paper is to analyse whether this provision can provide clear guidelines on how to articulate "equality by design", thus helping to limit some of the causes of algorithmic discrimination.

This paper is structured in two parts. The first part deals with the difficulties that the current legal framework on equality and non-discrimination, as well as its development through case law in the Court of Justice of the European Union (hereinafter, CJEU), present in providing clear guidelines that programmers can use to create non-discriminatory systems. The second part analyses article 10 of the proposal for an EU AI Act in order to determine whether it can provide some answers as to how equality by design should be articulated in AI systems.

1. The difficulty of articulating equality by design in the current regulatory and jurisprudential framework

In the context of AI, equality by design refers to the integration of the principle of equality in the process of developing AI systems with the purpose of ensuring that they do not

[1] For further analysis of the concepts and approaches to equality and non-discrimination, see Soriano Arnanz (2020), pp. 59-113.

generate discriminatory effects once they are deployed. As noted in the first section, one of the reasons why it is difficult to determine what parameters an AI system must meet to be considered non-discriminatory is the abstract nature of legal norms, which are designed to be adapted when applied to specific cases. As is evident, the difficulty of translating normative abstraction to the specificity required by computer code occurs not only in the area of discrimination but also in other areas, such as when incorporating data protection obligations into the system. However, the exact definition of what constitutes a discriminatory decision is particularly complicated for several reasons.

First of all, if we focus on the field of European law, this legal framework is not only characterised by the typical abstraction of the law, but by an added level of abstraction that results from the fact that these rules must subsequently be adapted to the context of each Member State. Thus, even if we examine the case law of the CJEU, we do not find a consistent set of criteria that could be translated into requirements to be considered in the programming of Artificial Intelligence systems. For example, the EU legal framework and case law have not set exact thresholds, for instance, on what percentage of women should be negatively affected by a decision or measure, in order to determine when a practice is to be considered discriminatory.

The European legal framework on equality and non-discrimination is established in article 21 of the Charter of Fundamental Rights of the EU (hereinafter, CFEU), which establishes the general clause prohibiting discrimination on the basis of an open list of certain specially-protected categories. Similarly, both the Treaty on the Functioning of the European Union (hereinafter TFEU) and the Treaty on European Union (hereinafter TEU) contain some precepts that provide for the generic protection of equality (articles 2 and 3 TEU and 8 TFEU) or in a more specific manner, referring for example to the protection and promotion of equality between women and men in the field of employment (articles 153 and 157 TFEU).

The generic prohibitions of discrimination contained in EU primary law, among which the clause of article 21 CFEU should be highlighted, are further developed in the Equality Directives. These Directives prohibit discrimination on grounds of race in employment, occupation, vocational training, various areas of social assistance, including social security and education, and access to goods and services;[2] discrimination on grounds of gender in self-employment,[3] employment, occupation, social security[4] and access to goods and services;[5] and discrimination on grounds of religion or belief, disability, age or sexual orientation in employment, occupation and vocational training.[6] All these rules provide a more specific definition of what constitutes discrimination and, in addition, establish two types of discrimination: direct and indirect.

In order to prove that an action is directly discriminatory, it is necessary to prove that a person is, has been or could be treated less favourably than another in a similar situation on the basis of one of the protected grounds. The possibilities of justifying a directly discriminatory measure by the defendant are very limited, since only objective and limited justifications are allowed, such as, for example, that the suspect category "constitutes a genuine and determining occupational requirement" – article 14(2) of the Directive on equality between women and men in employment.

Indirect discrimination occurs when an apparently neutral provision, criterion or practice is likely to place individuals that pertain to a protected group at a particular disadvantage compared to non-members of the group. Once the plaintiff proves a *prima facie* case of discrimination, the burden is on the defendant to prove that the apparently neutral provision, criterion or practice pursues a legitimate aim and that it is appropriate, necessary and proportionate to achieve that aim. For example, introducing benefits only for employees who work full-time is discriminatory against women because they are far more likely than men to take part-time jobs as a result of holding most caregiving responsibilities within families (Soriano Arnanz, 2021b, p. 115).

[2] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

[3] Directive 2010/41/EU of the European Parliament and of the Council of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity in a self-employed capacity and repealing Council Directive 86/613/EEC.

[4] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast).

[5] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services

[6] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.

In order to take these standards into account when programming non-discriminatory automated systems, we must know what is considered "less favourable treatment" or a "particular disadvantage". For example, programmers designing algorithms for recruiting employees should have a specific reference regarding the maximum percentage of a group that can be negatively affected by the system in order for it not to be considered discriminatory. For example, do we consider the system to be discriminatory if it systematically recommends men's CVs at a ratio of 70 to 30 with regard to women's? Or should the threshold be set at 60 to 40? Or is anything different from 50-50 to be considered discriminatory? If programmers have this exact reference, they can adjust systems before deploying them in order to ensure that they respect the principle and right to equality and non-discrimination. However, this is precisely where we encounter one of the main problems concerning the implementation of European equality standards in computer code, since neither European legislation nor European case law establishes a fixed threshold for considering an action to be discriminatory.

The ambiguity in determining which actions are considered discriminatory is an element that makes sense in the European context if we take into account the differences between the legal systems of the different EU Member States. Thus, the lack of precise determination of a threshold above which a practice is considered discriminatory allows for a more flexible application of the rules at the internal level of each State. However, this flexibility makes the work of those in charge of designing automated systems extremely difficult, as they do not have a clear reference to ensure that the system is not discriminatory.

Furthermore, another issue that hinders the translation of legal rules into computer code in the context of discrimination claims is the scarce use of statistics to measure discrimination claims at the EU level, as both national courts and the CJEU tend to rely more on traditional forms of evidence, such as common sense (Makkonen, 2007, p. 30, 34; Wachter *et al*., 2021, pp. 14-16).

The use of such more traditional forms of evidence generally involves making intuitive connections between the apparently neutral criterion and the discriminatory outcome. For example, in the United Kingdom, a ban on wearing turbans at work was found by the courts to be a clear case of indirect discrimination against ethnic Sikhs.[7] The problem that arises in the field of algorithmic discrimination is that it is not always easy to detect the relationship between the apparently neutral characteristic taken into account by the automated system and the membership of a disadvantaged group (Barocas & Selbst, 2018). For example, one of the criteria taken into consideration by the algorithm could be the colours that individuals prefer to purchase when buying clothes, and it may transpire that the colours chosen are associated with the race of individuals. The link between these two elements will clearly be hard to find and explain. Moreover, given the opacity that characterises algorithmic systems (Soriano Arnanz, 2021c, pp. 94-96), it is even more complicated to use traditional forms of evidence in cases of algorithmic discrimination, both indirect and direct, because we will not even know which criterion is causing the system to have a discriminatory impact.[8] This is why the prosecution of discrimination cases mediated by the use of artificial intelligence tools will lead to an increase in the use of statistical evidence.

It is therefore relevant to examine the criteria adopted by European courts and, in particular, the CJEU when accepting statistical evidence in cases of discrimination. At the European level, it has generally been required that the proportion of members of the disadvantaged group who are adversely affected by the indirectly discriminatory decision must be considerably high. Specifically, the CJEU considers that indirect discrimination exists when the apparently neutral measure "works to the disadvantage of far more" members of the disadvantaged group than non-members.[9] This expression was specified, among other decisions, in the Opinion of Advocate General Léger, issued in case C-317/93 Inge Nolte v. Landesversicherungsanstalt Hannover, in which he indicated that proving that 60% of the persons adversely affected by a measure belonged to a specially protected group was insufficient to consider such a measure as indirectly discriminatory. The

[7] Mandla (Sewa Singh) and another v. Dowell Lee and others [1983] 2 AC 548.

[8] It is worth noting, in relation to proof of direct algorithmic discrimination, that without full access to the system, it will be extremely difficult to prove direct discrimination, unless the system clearly treats all members of the disadvantaged group in a less favourable manner.

[9] See, for instance, CJEU Judgment 20 March 2011, C-123/10, Brachner (paragraph 56); 22 November 2011, C-385/11, Elbal Moreno (paragraph 29); 18 March 2014, C-167/12, C.D. v. S.T (paragraph 48); and 14 April 2015, C-527/13, Lourdes Cachaldora Fernández v. INSS (paragraph 28).

statistical threshold to consider and apparently neutral measure as discriminatory has thus been generally placed at 80%. That is, if this criterion is applied, at least 80% of the individuals negatively affected by the measure must belong to the disadvantaged group (Wachter, 2020, p. 45).

However, it is highly relevant to bear in mind that this percentage is by no means a fixed figure and that, in fact, the CJEU has recognised that it could be accepted that a smaller proportion of individuals adversely affected by the measure belonged to the disadvantaged group "if the statistical evidence revealed a lesser but persistent and relatively constant disparity over a long period".[10] Considering this ruling along with the fact that AI systems are used to process many people, it would be possible to lower the threshold for considering that a decision made by or with the help of an AI system is discriminatory.

In any case, it is obvious that there is no fixed criterion for determining what should be considered an indirectly discriminatory decision, which makes it very difficult for programmers to create non-discriminatory systems, as they have no clear rule to follow.

It is worth highlighting that in the US, unlike in the EU, a metric threshold is established, at least in the field of employment, above which the measure adopted is considered to be discriminatory. This rule, known as "the four-fifths rule",[11] requires that the proportion of recruits of any race or sex must not be less than four-fifths or 80% of the selected individuals belonging to the group with the highest selection rate. These percentages are calculated based on the number of people from each group who applied for the job. Thus, if 100% of the men applying for a job were selected, the percentage of women hired should be 80% of those who applied (U.S. Equal Employment Opportunity Commission, 1979).

While it is true that the four-fifths rule does not constitute a fixed and immovable threshold, as the situation in each specific case must also be analysed (Barocas & Selbst, 2016, p. 702), what is certain is that it at least provides a clear criterion that can be taken into account by female and male programmers.

2. Equality by design in the proposed EU Artificial Intelligence Regulation

2.1. *Ex ante* control solutions

One of the main shortcomings of the existing legal framework for the protection of equality and non-discrimination is that the prohibitions to discriminate are mechanisms that operate *ex post*, that is, after the discriminatory action has already taken place. While it is true that there is an obligation not to discriminate, which serves as a starting point for articulating equality by design in artificial intelligence systems, the fact is that there are no *ex ante* regulatory control mechanisms to ensure that AI systems are not discriminatory.

2.2. Requirements to be met by training, validation and test data for high-risk AI systems

Automated systems are trained with data related to the phenomenon they seek to predict. For example, a system designed to determine the most suitable candidates for a job can be trained with historical data regarding a company's recruitment processes.

Once the system has been trained, its performance will be evaluated with validation data. This data is used to detect the level of accuracy of the system and, in general, to verify that the system adequately measures and predicts the aspect of social reality it is supposed to process and evaluate. If we are designing a system for generating profiles of possible perpetrators of a homicide, the validation data set will contain information related to homicides that have already been solved. The part of the information obtained during the investigation will be entered into the system without indicating the characteristics of the perpetrator in order to check whether the prediction made by the system corresponds to reality.

Finally, as stated in article 3.31 of the proposed AI Regulation, test data are "data used for providing an independent evaluation of the trained and validated AI system in

[10] CJEU Judgement, 9 February 1999, C-167/97, Regina v. Secretary of state for Employment, ex parte: Nicole Seymour-Smith and Laura Perez (paragraph 61).

[11] This rule was first published in 1978 in the "Uniform Guidelines on employee selection procedures", section § 1607.4.D Title 29 US Code of Federal Regulations.

order to confirm the expected performance of that system before its placing on the market or putting into service".

The origin of algorithmic discrimination can often be found in the datasets used to train the system (Barocas & Selbst, 2016; Hacker, 2018). Considering that society has historically been built on structures of discrimination that have placed, and still place, certain population groups in positions of disadvantage or subordination, when an automated system is trained with data from the "real world", it is easy for that information to be contaminated by the indicated structures of discrimination, thus leading the system to internalise the disadvantaged position in which society places certain groups. Similarly, if the validation and test data used to train the system also reflect these historical structures of discrimination, the system will confirm that the biases it contains are, in fact, correct.

Therefore, establishing requirements with which the system's training, validation and test data must comply is not only useful but also essential as a mechanism for preventing algorithmic discrimination. This is recognised in Recital 44 of the proposed Regulation by stating that:

> "High data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are u herramientas sed, with a view to ensure that the high-risk AI system performs as intended and safely and it does not become the source of discrimination prohibited by Union law."

In this regard, article 10 of the proposed AI Regulation sets out a number of requirements to be met by the training, validation and test data of high-risk AI systems, as well as by the people or organisations in charge of collecting such data and processing them. The following pages discuss how the mandates of article 10(2) to (4) address different issues that, if not considered in the initial system design, data collection and processing phases, can lead to bias and discriminatory results.

2.2.1. Formulation of assumptions and selection of characteristics

The second paragraph of article 10 states that "training, validation and testing data sets shall be subject to appropriate data governance and management practices", and goes on to list those aspects on which such practices should focus.

Among other elements, these good practices should focus on "the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent" (Art. 10.2.d) and the "prior assessment of the availability, quantity and suitability of the data sets that are needed" (Art. 10.2.e). Similarly, article 10(3) states that "training, validation and testing data sets shall be relevant, representative [...]."

Finally, with regard to the elements relevant to the analysis carried out in this section, article 10(4) states the following:

> "Training, validation and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used."

The transcribed requirements can be easily summarised in a single word: appropriateness. The purpose of these mandates is to ensure that the data are suitable for measuring the aspect of social reality or human behaviour that the system is designed to analyse or predict. The requirements detailed above are particularly relevant in relation to two of the initial phases of the design and training process of an algorithmic system: problem specification (and formulation of assumptions) and feature selection.

Problem specification is the definition of the objective to be pursued. For example, the objective could be to determine the probability that each applicant for a mortgage loan will default on the repayment conditions of the loan. This objective is divided into different possible outcomes or assumptions that must be adequate to predict the behaviour measured. For instance, in the case of predicting default in the repayment of a mortgage loan, it is probably more suitable to express the results with sequential numerical values (from 0 to 100) than with fixed categories such as the classification into high, medium and low probability of default. This is because the conditions under which the loan will be granted will depend on the category in which each person is classified. Therefore, if three broad categories are established, it is likely that all individuals with a 67 to 100 probability of default will be assigned to the group with the highest probability of not being able to repay the loan, and the loan granting (or denial) conditions will be identical for all of them, despite the significant differences that will be found within the group itself. This is why the formulation of the assumptions is of great importance to ensure that decisions made based on the system's predictions are appropriate and to avoid treating people in different situations in an equally detrimental manner.

The selection of the characteristics or variables (data) that are taken into account when measuring the aspect of social reality or human behaviour that the system intends to predict is also of enormous relevance, because it is one of the main ways in which situations of indirect discrimination can be generated by AI systems. In this sense, it is easy to choose variables that, despite being apparently neutral, in fact correlate to belonging or not belonging to a disadvantaged group. For example, the valuation of postal codes in decision-making involves the valuation of an apparently neutral criterion, but one which in the USA has been shown to be closely linked to belonging to one racial or ethnic group or another (Hunt, 2005).

2.2.2. Detection of gaps, biases and errors

Section 10(2)(f) states that examination of the data should be conducted "in view possible biases". Subsequently, paragraph (g) states that "any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed" should be identified.

Artificial intelligence systems are often not completely objective. One of the main reasons why these systems may be biased is as a consequence of having learned from a database that under- or over-represents disadvantaged groups, so that the decisions made by the system end up harming them (Žliobaitė, 2015). If the training data of a system employed for personnel selection in a company are biased so that women are underrepresented, the system will learn to eliminate or give lower scores to female job applicants than to male candidates. On the other hand, if a system used to predict whether a person has a high probability of defaulting on loan repayment terms is trained with a database in which people of a certain ethnicity or minority race are over-represented, the system will attribute a higher risk of default to them.

Errors or gaps in the databases are also common, especially considering that designers and operators of algorithmic systems tend to use cheaper databases (Barocas & Selbst, 2016, p. 689). These databases are very effective in creating algorithmic systems since, despite the possible errors or gaps they may contain, they remain enormously accurate. However, these errors and gaps tend to be found in the data regarding members to disadvantaged groups, leading to the system making more errors with respect to these groups once it has been implemented (Kim, 2015, pp. 885-886).

Also in relation to these issues, article 10(3) states that "training, validation and testing data sets shall be relevant, representative, free of errors and complete".

2.2.3. Labelling

Article 10(2)(c) of the proposed AI Regulation states that "relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation".

In supervised learning environments, the data with which the system is trained are labelled, grouped and classified so that the algorithms learn what characteristics to look for in the people they analyse once they are put into operation and, based on the characteristics detected, how each person should be classified. In unsupervised learning systems, on the other hand, it is the system itself that must autonomously identify the existing relationships between the different categories of data it is fed (Gerards & Xenidis, 2021, pp. 34-35). Consequently, the labelling and classification phases are part of the data processing involved in the programming of supervised learning systems.

For example, a system used to predict the creditworthiness of individuals applying for a bank loan will be trained to detect certain characteristics in applicants that are relevant to the granting of a loan, such as income level and savings. However, the system can also be trained to consider other characteristics that are less relevant in determining an individual's creditworthiness, such as the type of music they listen to, and which may contribute to the introduction of biases into the system that favour the members of historically advantaged groups.

2.3. Effectiveness and shortcomings of article 10 as a mechanism to ensure equality by design in AI systems

As the brief analysis of some of the sections contained in article 10 of the proposed AI Regulation conveys, these rules are still characterised by containing general mandates that are difficult to specify for people who design AI systems. Thus, although the provision refers to the various moments in the design of algorithmic systems at which decisions can be made that may result in the system perpetuating the historical structures of discrimination that underlie society, the fact is that it does not specify what these "relevant data preparation processing operations" for labelling are, nor what is considered a biased database, among other issues.

However, we should still acknowledge the positive aspects of this provision insofar as it focuses on highlighting those phases of the design process of AI systems to which special attention should be paid, which itself enhances the chances of detecting possible elements that may lead to discriminatory results and preventing these, especially if we consider that the people in charge of designing automated systems are often unaware of the risks to the rights to equality and non-discrimination that they may generate. Moreover, we must also finally highlight the importance of article 10.5, as it refers to the possibility of processing special categories of data, which, for instance, include race, religious beliefs and biometric data, when the purpose is to monitor, detect and correct possible biases in high-risk AI systems. Hence, this specific section not only focuses on the design of algorithmic systems from the perspective of equality, but also provides a mechanism that can contribute to the detection of algorithmic discrimination.

Article 10.5 of the proposal for an EU AI Act is also particularly relevant because it establishes an exception to the prohibition of processing special categories of data set, amongst other rules, by article 9 of the GDPR, which refers to the processing of information that, to a large extent, can be identified with the suspect categories of producing discrimination.

Article 9 GDPR has been criticised on the grounds that the impossibility of processing these data categories may increase the possibility of hiding discriminatory instructions in algorithms (Soriano Arnanz, 2020, pp. 395-404; Žliobaitė & Custers, 2016, p. 198). This is why introducing an exemption to the prohibition when what is intended is precisely to detect these potential biases is very useful, as this provision serves as the necessary legal basis for developing tools aimed at detecting and preventing the perpetuation of inequality mediated by the use of AI systems.

Conclusions

Throughout the previous pages, the lack of specific legal norms, in particular, in the area of equality and non-discrimination, has been presented as one of the main obstacles to be overcome if we want AI systems to respect the legal system from the moment they are created. The legal framework on equality and non-discrimination and case law developed at the European level do not establish sufficiently precise mandates that programmers can translate into computer code.

This is why it is necessary to pass specific rules, such as the proposed regulation on Artificial Intelligence, that regulate AI decision-making. From the analysis of article 10 of said proposed regulation, it is once again possible to identify a lack of specification in the rules that refer to the data used in the design and creation of automated systems. However, the simple fact that the aspects of the data selection and processing phases that can generate biases are pointed out, as well as having highlighted these risks, is already a huge step forward with respect to other existing rules and could help people in charge of designing AI systems to be aware of and control the possible appearance of biases and even create tools to ensure an ex post control even after AI systems are deployed.

Acknowledgments

This paper has been carried out in the framework of the SHINE (Sharing Economy and Inequalities across Europe) Jean Monnet Network, with the support of the European Union and the Algorithmical Law (PROMETEU/2021/009) research project, funded by the Generalitat Valenciana.

References

BAROCAS, S.; SELBST, A. D. (2016). "Big data's disparate impact". *California Law Review*, vol. 104, no. 3, pp. 671-732. DOI: https://doi.org/10.2139/ssrn.2477899

BAROCAS, S.; SELBST, A. D. (2018). "The intuitive appeal of explainable machines". *Fordham Law Review*, vol. 87, no. 3, pp. 1085-1139. DOI: https://doi.org/10.2139/ssrn.3126971

BENT, J. R. (2020). "Is algorithmic affirmative action legal?". *The Georgetown Law Journal*, vol. 108, pp. 803-853.

BERK, R.; HEIDARI, H.; JABBARI, S.; KEARNS, M.; ROTH, A. (2018). "Fairness in criminal justice risk assessments: the state of the art". *Sociological Methods and Research*, vol 50, no. 1, pp. 1-24. DOI: https://doi.org/10.1177/0049124118782533

CERRILLO I MARTÍNEZ, A. (2020). "El impacto de la inteligencia artificial en el derecho administrativo ¿nuevos conceptos para nuevas realidades técnicas?". *Revista General de Derecho Administrativo*, no. 50.

CHOULDECHOVA, A. (2016). "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments". *arXiv* [online]. [Accessed: 6 September 2022]. DOI: https://doi.org/10.48550/arXiv.1610.07524

CORBETT-DAVIES, S.; PIERSON, E.; GOEL, S. (2015, October). "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear". *The Washington Post* [online]. [Accessed: 6 September 2022]. Available at: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?noredirect=on

CORBETT-DAVIES, S.; GOEL, S. (2018). "The measure and mismeasure of fairness: a critical review of fair machine learning". *ArXiv* [online]. [Accessed: 6 September 2022]. DOI: https://doi.org/10.48550/arXiv.1808.00023

FRIEDLER, S. A.; SCHEIDEGGER, C. E.; VENKATASUBRAMANIAN, S.; CHOUDHARY, S.; HAMILTON, E. P.; ROTH, D. (2018). "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning". *Proceedings of the Conference on Fairness, Accountability, and Transparency* [online]. [Accessed: 6 September 2022]. DOI: https://doi.org/10.1145/3287560.3287589

GERARDS, J.; XENIDIS, R. (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: Publications Office of the European Union.

HACKER, P. (2018). "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law". *Common Market Law Review*, vol. 55, no. 4, pp.1143-1186. DOI: https://doi.org/10.54648/COLA2018095

HUERGO LORA, A. (2020). "Una aproximación a los algoritmos desde el Derecho administrativo". In: HUERGO LORA. A. (dir.) and Díaz González, G.M. (coord.). *La Regulación de los Algoritmos*, pp. 23-87. Cizur Menor: Aranzadi.

HUNT, B. (2005). "Redlining". *Encyclopedia of Chicago*, 2005 [online]. [Accessed: 6 September 2022]. Available at: http://www.encyclopedia.chicagohistory.org/

KIM, P. T. (2017). "Data-driven discrimination at work". *William & Mary Law Review*, vol. 58, pp. 857-936.

LESSIG, L. (2006). *Code: Version 2.0*. New York: Basic books.

MAKKONEN, T. (2007). *Measuring discrimination: data collection and EU equality law*. Luxembourg: Office for Official Publications of the European Communities.

PLEISS, G.; RAGHAVAN, M.; WU, F.; KLEINBERG, J.; WEINBERGER, K. Q. (2017). "On Fairness and Calibration". *Advances in Neural Information Processing Systems*. [online]. [Accessed: 6 September 2022]. Available at: https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

RENAN BARZILAY, A.; BEN-DAVID, A. (2017). "Platform inequality: gender in the gig economy". *Seton Hall Law Review*, vol. 47, no. 393, pp. 393-431. DOI: https://doi.org/10.2139/ssrn.2995906

SORIANO ARNANZ, A. (2020). *Posibilidades actuales y futuras para la regulación de la discriminación producida por algoritmos*. Doctoral tesis [online]. [Accessed: 6 September 2022]. Available at: https://roderic.uv.es/handle/10550/77050

SORIANO ARNANZ, A. (2021a). "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*, vol. 8, no. 1.

SORIANO ARNANZ, A. (2021b). "La situación de las mujeres en el empleo público: análisis y propuestas". *IgualdadES*, no. 4, pp. 87-121. DOI: https://doi.org/10.18042/cepc/IgdES.4.03

SORIANO ARNANZ, A. (2021c). "Decisiones automatizadas: problemas y soluciones jurídicas. Más allá de la protección de datos". *Revista de Derecho Público: Teoría y Método*, vol. 3, pp. 85-127. DOI: https://doi.org/10.37417/RPD/vol_3_2021_535

US Equal Employment Opportunity Commission. (1979). "Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures". *US Equal Employment Opportunity Commission* [online]. [Accessed: 6 September 2022]. Available at: https://www.eeoc.gov/

VALERO TORRIJOS, J. (2020). "The legal guarantees of artificial intelligence in administrative activity: reflections and contributions from the viewpoint of Spanish administrative law and good administration requirements". *European Review of Digital Administration & Law*, vol. 1, no. 1-2, pp. 55-62.

WACHTER, S. (2020). "Affinity Profiling and Discrimination by Association in Online Behavioural Advertising". *Berkeley Technology Law Journal*, vol. 35, no. 2. DOI: https://doi.org/10.2139/ssrn.3388639

WACHTER, S.; MITTELSTADT, B.; RUSSELL, C. (2021). "Why fairnesss cannot be automated: bridging the gap between EU non-discrimination law and AI". *Computer Law & Security Review*, vol. 41. DOI: https://doi.org/10.1016/j.clsr.2021.105567

ŽLIOBAITĖ, I. (2015). "A survey on measuring indirect discrimination in machine learning". *arXiv* [online]. [Accessed: 6 September 2022]. DOI: https://doi.org/10.48550/arXiv.1511.00148

ŽLIOBAITĖ, I.; CUSTERS, B. (2016). "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models". *Artificial Intelligence & Law*, vol. 24, no. 2, pp. 183-201. DOI: https://doi.org/10.1007/s10506-016-9182-5.

Recommended citation

SORIANO ARNANZ, Alba (2022). "Creating non-discriminatory Artificial Intelligence systems: balancing the tensions between code granularity and the general nature of legal rules". *IDP. Internet, Law and Politics E-Journal*. No. 38. UOC [Accessed: dd/mm/aa]

About the author

Alba Soriano Arnanz
Universitat de València
alba.soriano@uv.es

Assistant Professor of Administrative Law, PhD at the University of Valencia. Graduate in Law and Political Sciences and Public Administration from the University of Valencia (2015), Master of International Political Economy from the London School of Economics and Political Science (2016), Master of Law from the UOC (2018) and Doctor of Law from the University of Valencia with his doctoral thesis *Current and future possibilities for the regulation of discrimination produced by algorithms* (2021), led by Professor Andrés Boix Palop. She is the author of various publications on the subject of personal data protection and algorithmic discrimination, among which the book *Data protection for the prevention of algorithmic discrimination* (2021) edited by Thomson Reuters – Aranzadi and the articles "Automated decision-making and discrimination: general approach and proposals", *Revista General de Derecho Administrativo*, no. 56, 2021 and "Automated Decisions – Legal Problems and Solutions. Beyond Data Protection," *Revista de Derecho Público: Teoría y Método*, no. 1.3, 2021 stand out. She has also dedicated some of her research to improving recruitment systems and the situation of women in public employment.

Universitat Oberta de Catalunya