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

https://idp.uoc.edu

Dossier "Europe facing the digital challenge: obstacles and solutions"

Is the EU human rights legal framework able to cope with discriminatory AI?

Pablo Martínez-Ramil
Palacký University Olomouc

Date of submission: May 2021
Accepted in: September 2021
Published in: December 2021

Abstract

The challenges introduced by AI for the EU anti-discrimination legal framework have been a widely discussed topic among the doctrine. In the light of the 20th anniversary of the EU Charter of Fundamental Rights, the Commission released a regulatory proposal to tackle AI. This paper seeks to determine whether the proposal successfully addresses the existent pitfalls of the EU framework. First, this paper explores the functioning of AI systems that employ machine learning techniques and determines how discrimination takes place. Second, the article examines intellectual property rights as one of the main barriers for accountability and redressal of violations committed by an AI system. Third, the state of the discussion concerning the pitfalls of the existent EU approach towards non-discrimination is addressed. The available academic literature suggests that discriminatory outputs produced by an AI will amount to indirect discrimination in most scenarios. In this sense, cases of indirect proxy discrimination will likely pass the proportionality test, therefore justifying the discriminatory output. The last section of this article studies the Commission's regulatory proposal. Although the document seems to effectively tackle discrimination caused by biased training data sets, this paper concludes that intellectual property rights and proxy discrimination still constitute significant barriers for the enforcement of anti-discrimination law.

Keywords

human rights; Charter of Fundamental Rights of the European Union; artificial antelligence; AI; algorithms; trade secrecy law; non-discrimination

¿Es el marco legal de derechos humanos de la UE capaz de hacer frente a la IA discriminatoria?

Resumen

Los desafíos introducidos por IA para el marco legal de la UE contra la discriminación han sido un tema ampliamente discutido entre la doctrina. A la luz del vigésimo aniversario de la Carta de los Derechos Fundamentales de la UE, la Comisión publicó una propuesta reglamentaria para abordar la IA. Este documento busca determinar si la propuesta aborda con éxito los escollos existentes del marco de la UE. En primer lugar, este documento explora el funcionamiento de los sistemas de inteligencia artificial que emplean técnicas de aprendizaje automático y determina cómo se produce la discriminación. En segundo lugar, el artículo examina los derechos de propiedad intelectual como una de las principales barreras para la rendición de cuentas y la reparación de las violaciones cometidas por un sistema de inteligencia artificial. En tercer lugar, se aborda el estado del debate sobre los escollos del actual enfoque de la UE hacia la no discriminación. La literatura académica disponible sugiere que los resultados discriminatorios producidos por una IA equivaldrán a discriminación indirecta en la mayoría de los escenarios. En este sentido, los casos de discriminación indirecta por proxy probablemente pasarán la prueba de proporcionalidad, lo que justificará el resultado discriminatorio. La última sección de este artículo estudia la propuesta reguladora de la Comisión. Aunque el documento parece abordar de manera efectiva la discriminación causada por conjuntos de datos de capacitación sesgados, este documento concluye que los derechos de propiedad intelectual y la discriminación por poder aún constituyen barreras significativas para la aplicación de la ley contra la discriminación.

Palabras clave

derechos humanos; Carta de los Derechos Fundamentales de la Unión Europea; Inteligencia Artificial; IA; algoritmos; ley de secreto comercial; no discriminación

Introduction

Modern societies are far away from being perfect, dealing for years with problems such as discrimination based on gender or race. Due to the fact that Artificial Intelligence (hereinafter AI) technologies "learn" how to act from the available (and biased) data, more than often these technologies reproduce in their results the existing problems in our societies (Andersen, 2018, p. 12). This raised the alarms in the legal realm, where many voices warned about the AI capacity to impact negatively in a variety of human rights (Aizenberg & Van den Hoven, 2020, pp. 1-2).

Several cases made it to the headlines over the last few years. In the US justice system, an AI technology named COMPAS was employed to determine the convicts' risk of reoffending when deciding penalties. The AI classified 45% of those African-American convicts who ultimately did not reoffend as "high risk," as compared to just 23% for Caucasians in a similar situation (Raso & others, 2018, p. 23). Likewise, the automatized Amazon's recruiting system proved to show clear discriminatory results against women. Because the AI was "fed" with the recruitment data from the previous 10 years and since the majority of the hired workers in that period were men (a reflection of male dominance in the tech industry), the AI acted in consonance (Dastin, 2018). Some emerging issues have already been contested in the judicial sphere. Last year, the District Court of The Hague ruled that the right to privacy prevailed over the legality of the System Risk Indication (SYRI), a system that allowed the government to process large amounts of data collected by public authorities to identify those most likely to commit benefits fraud (Henley & Booth, 2020).

It is imperative to note at an early stage that the origins and development of International Human Rights Law (hereinafter IHRL) took place in entirely analogic times. Therefore, it has often struggled with technological disruptions. As a technology, AI constitutes a great one. It can be incorporated in almost any kind of human activity, meaning that "the breadth of regulatory issues raised by AI span the spectrum of human activities" (Liu & others, 2020, p. 7). That is why balancing AI with a strong legal framework mindful of human rights will become increasingly important in the near future.

The European Union (hereinafter EU) was fully aware of that. Hence, after dealing with Brexit, the refugee crisis and while facing issues such as the state of the rule of law in Poland and Hungary, the time has come to address the challenges introduced by AI. Bearing in mind the 20th anniversary of the Charter of Fundamental Rights (hereinafter the Charter), the Commission has released a regulatory proposal for establishing an AI legal framework (hereinafter the AI Act). In line with the above, one of its main objectives is to ensure that AI systems comply with the existing law on fundamental rights and Union values.

Although it is evident that AI has become the talk of the town for quite a while now, no consensus has been reached yet concerning the content of its definition. This paper will refer here to the one provided by the UNESCO' expert group, which defined AI as "technological systems which have the capacity to process information in a way that resembles intelligent behavior, and typically includes aspects of reasoning, learning, perception, prediction, planning or control" (Ad Hoc Expert Group for the preparation of a draft text of a recommendation on the ethics of artificial intelligence, 2020, p. 4).

This research is motivated by the following question: "is the established EU human rights legal framework able to cope with the challenges for non-discrimination introduced by AI?"

First, several features of AI and machine learning systems will be addressed. Their functioning as well as some basic concepts will be examined. The analysis will continue addressing legal issues that test the suitability of the EU approach towards non-discrimination as the ideal framework to face the upcoming AI challenges. Questions such as those concerning trade secrecy law and its implications for accountability and redressal will be examined. Then, the Commission proposal will be briefly discussed considering the previously studied pitfalls. An answer to the research question will conclude this article.

1. Discriminatory AI. What, why and how?

At first, an AI system was a software that interpreted data using an algorithm, a mathematical formula employed to produce the required results. Therefore, any outcome perceived as unfair could be explained after a close examination of both the data and the algorithm (Andersen, 2018, pp. 9-12).

However, this approach has become outdated. The vast majority of contemporary AI systems do not simply constitute a given set of rules that analyze data. They are built instead upon a range of techniques that differ on their functioning. Therefore, a detailed analysis covering the wide (and growing) variety of approaches and AI techniques would justify more than one peer-reviewed journal submission. For this reason, the scope of this section will cover the techniques commonly referred as machine learning (hereinafter ML), the most widespread area of AI being practically applied (Furtwangen University). In particular, it will analyze how discrimination occurs within supervised and unsupervised ML environments.

In a few words, ML works as an umbrella term that addresses those algorithmic models that allow the AI to learn "by example" (Hacker, 2018, p. 5). Supervised and unsupervised ML systems learn how to produce outputs after being trained with large amounts of data. The main difference between both is that the former uses labelled data sets. "Labelling" in ML refers to the process of adding informative labels to raw data.[1] This procedure is usually either done by the AI developing team or by an external contractor. Therefore, labelled data sets are usually costlier and contain smaller amounts of values (Dilmegani, 2021).

In a preliminary stage, the (unlabelled/labelled) data is divided into two sets: training data and validating data. The former is used to teach the right outcomes to the system (through inputs and outputs) and the later serves as a control test (only inputs). Taking Amazon's recruitment system as an example, the training data set would encompass both the data contained in the professional CVs previously processed by the company (inputs) and the final decision on whether the applicant got the job or not (output). Once the model is exposed to this information, it creates a formula that explains the relationship between inputs and outputs (to rephrase it, why some applicants were hired and some not). In supervised ML contexts, the developers might "influence" this resultant formula by assigning a determined weight to certain values or variables (for instance, using a "decision tree").[2] In the highlighted example, the developers could have established that "previous labour experience" holds more significance than "education" when determining whether an applicant must be hired or not.

Once the system has developed a model, it needs to be "tested" through the validating data set. The formula, being only exposed to the aforementioned CVs, has to determine the right answer (to hire or not to hire). If it provides the right outcomes, the system is tested again using data from the real world (usually referred as testing data set) (Hacker, 2018, p. 5). Only after succeeding at the second test the system might be deployed for its use.

Before analysing the origins of discriminatory AI, it is necessary to draft some notes on deep learning, a ML sub-technique. The above-exposed ML procedures comprise three layers. The (i) inputs are interpreted through a (ii) model (hidden layer) that provides an (iii) output. This is often referred as "neural network", a set of algorithms interpreting the relationship between inputs and outputs. In "traditional" supervised and unsupervised ML, the neural network is composed of only three layers. In comparison, deep learning is a subfield of ML where a greater number of hidden layers exists, with every layer being a tool to transform the input data into a slightly more abstract representation of itself to infer correlations. As a consequence, the explanatory algorithm that determines the outputs shows a high level of complexity[3] (the more layers, the more complexity) with outcomes that can become untraceable to the human eye. In other words, in deep learning environments, even the developers might not be able to "understand" the reasoning behind a certain output.

[1] See, in this regard, AMAZON. *What is data labeling for machine learning?* [online]. Available at: https://aws.amazon.com/es/sagemaker/groundtruth/what-is-data-labeling/ [Accessed: 13 August 2021].

[2] See, in this regard, XORIANT (2017, September). *Decision Trees for Classification: A Machine Learning Algorithm* [blog]. Available at: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html [Accessed: 13 August 2021].

[3] See, in this regard, KOCHLING, A.; WEHNER, M. C. (2020). "Discriminated by an algorithm: a systemic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development". *Business Research,* vol. 13, pp. 795–848.

Based on Hacker's approach[4] this paper proposes the following classification to determine how discriminatory outputs are generated within this scheme:

- Biased training data. If the training data is biased, the model developed afterwards will produce biased results. Three main subtypes were identified.
  - Inadequate construction of training data sets. If a set is incomplete (missing certain values), not representative enough (a potential group of users is not adequately represented in the data set) or if it contains errors (false information in the inputs or outputs), the accuracy of the resultant algorithm might be compromised, increasing its discriminatory potential. An (in)famous illustrative incident took place when Google's AI mislabelled black people as gorillas. Apparently, the AI lacked training for identifying black peoples' faces (Zhang, 2015).
  - Incorrect labelling. In supervised ML contexts, some decisions on the AI system' design might lead to discriminatory outputs. For instance, a wrong "weight" assignment to certain variables within a decision tree or errors in the labelling procedure will likely compromise the accuracy of the resultant algorithm.
  - Historical bias of the data. Often, data sets reproduce the existent biases in our societies, leading consequently to biased outcomes. The above-mentioned case of Amazon's recruitment AI exemplifies this.

- Proxy discrimination. The academia widely addressed this type of discrimination, being the one raising most legal questions. As it was exposed, ML systems analyze training data sets looking for variables to explain correlations between outputs and inputs. If a training data set is infused with historical racial biases, the system will use the variable "race" to explain the correlations. However, due to the fact that ML systems develop their own predictive models from observation, an apparently unbiased training data set is also able to produce discriminatory results. This greatly hinders the detection of causes of a discriminatory algorithm (Hacker, 2018, pp. 5-6). For instance, an AI system that measures credit worthiness is being developed. Because of historical reasons, racial minorities have been systematically discriminated, facing, among other problems, lower incomes and higher levels of financial instability. In general, they would "score" lower values in the variables of a training data set. The developers, being aware of it, decide to remove the variable "race" from the training data set. And yet, the outputs turned out to be discriminatory. Why? Here, the AI might have developed a model using the variable "postal code" to explain the relationship between inputs and outputs. Certain levels of segregation still exist nowadays (Lisa, 2019), what often makes racial minorities share postal code. Therefore, even if the model did not rely on race as a variable to discriminate, the consequences remained unaltered. This is the classical textbook case of what can be defined as proxy discrimination. Proxy discrimination can occur in both supervised and unsupervised learning environments. It is very hard to predict and, in deep learning contexts, very hard to detect.

As some voices have pointed out, a designing-against-discrimination solution would require the developers to have a deep comprehension of historical and social reasons that explain discrimination. In addition, AI creators should pay special attention to technical aspects of the design (such as the representativity of the data collected, potential discriminatory consequences derived from the choice of variables or the definition of class labels) that might lead to the development of uncategorized types of discrimination (Aizenberg & Van den Hoven, 2020, p. 3). Moreover, even if an AI is carefully designed, the available evidence establishes that strategies for reducing bias during the development process often reduce the predictive potential of the AI. At the same time, the economical investment attached to the construction of new unbiased databases might not be bearable for some companies (Hacker, 2018, p. 8).

[4] For a more detailed analysis, see HACKER, P. (2018). "Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU Law". In: *Common Market Law Review*, vol. 55, no. 4, pp. 3-8.

2. Potential gaps of the EU Human Rights legal framework or how to deal with discriminatory AI

2.1. IP rights, accountability and redressal

The charter Art. 47 enshrines the right to an effective remedy for those whose rights have been violated. However, as it has been demonstrated, the technical characteristics of ai systems involve several issues that, at least, hinder the implementation of this provision in several ways.

To perceive and demonstrate that a violation has been committed by an AI system, it is necessary to have (I) a contextual knowledge of the technology and in some cases (II) access to the code. The question of transparency is rather problematic and full of cornerstones. On the one hand, because even with access to the code there is no guarantee that the human eye can trace the origins of a discriminatory output (particularly in deep learning environments). On the other hand, because the code might be protected by intellectual property (hereinafter IP) rights. This deserves a closer look.

A well-developed AI system constitutes a clear market advantage for the developing company. The AI guidelines published by the European Patent Office consider computational models and algorithms to be of mathematical nature (European Patent Office), not being therefore patentable under the applicable law.[5] Instead, they fall under the protection of trade secrecy law.

The EU Trade Secrets Directive defines "trade secret" as information that (I) is secret, (II) has commercial value and (III) whose secrecy has been object of protection.[6] This protection might be limited for reasons of public interest or in exercise of the right to freedom of expression.[7] However, the directive does not provide any sort of guideline to support the authorities when addressing a conflict between the public interest and the protection granted by trade secrecy (Huseinzade, 2021).

Hence, these legal conflicts have been approached differently at the national and European level. In *Microsoft*, the Court recognized that IP rights are not undefeatable or otherwise exceptions would never apply.[8] This was the prevailing view in *Google Shopping*, where Google was forced to disclose protected information for the Commission' investigative duties.

Some scholars have put the focus on the General Data Protection Regulation Art. 22. It grants the data subject a right "not to be subject to a decision based solely on automated processing (…) which produces legal effects."[9] This provision was criticized given its difficult enforceability. In practical terms, the only safeguard that introduces is the right to contest an automatized decision before a human agent, but it does not create a "right to explanation", understood as the right to be given an explanation for an output produced by an AI system (Wachter & others, 2017, pp. 79-81). Academics have concluded that, to this date, trade secrecy law still constitutes a significant barrier (Wachter & others, 2017, p. 89).

2.2. The pitfalls of the EU approach towards non-discrimination

The Charter Art. 21 enshrines the prohibition of discrimination.[10] The Court of Justice has systematically applied this provision in horizontal relations. Moreover, it has interpreted it widely, covering types of discrimination that arguably were not strictly part of the constitutional systems of the member states (see, in this regard, Mangold, Kücükdeveci and Egenberger). Secondary law has given expression to several dimensions of this prohibition (such as gender, race or religion).

The Commission's White Paper on Artificial Intelligence, a document that outlined in 2020 the AI policy options and

[5] Art. 52(2)(a). Convention on the Grant of European Patents of 5 October 1973.
[6] Art. 2 (1). EU Directive (EU) 2016/943.
[7] Arts. 1(2) and 5. Directive (EU) 2016/943.
[8] Case T-201/04 of 17 September 2007. *Microsoft Corp. v Commission of the European Communities*, para. 690.
[9] Art. 22 (1). Regulation (EU) 2016/679 (General Data Protection Regulation).
[10] EUROPEAN PARLIAMENT (2000). *Charter of fundamental rights of the European Union*, Art. 21(1). Luxembourg: Office for Official Publications of the European Communities.

set the path for the elaboration of the AI Act, establishes that the EU regulative framework remains applicable "irrespective of the involvement of AI" (European Commission, 2020). However, the differences in the scope of the directives generate different levels of protection. While in employment matters the protected grounds are religion, disability, age and sexual orientation, in goods and services only race and gender are covered by EU law. Part of the academia (Hacker, 2018, pp. 9-10) defends that a possible solution could come by the hand of the long-standing Commission proposal to extend the scope of protection (European Commission, 2008).

The directives distinguish two types of discrimination: direct and indirect. The former would be rather rare in ML contexts. It occurs when a person is treated less favourably than another – in a comparable situation – based on a protected ground (such as sexual orientation, gender, race…).[11] Within the AI realm, direct discrimination could allegedly take place due to an inadequate construction of training data sets or wrongful labelling procedures. For example, if an AI developer had deliberately assigned lower values to a label or variable expressing a protected ground (such as race or gender), the conduct would amount to direct discrimination. It should be noticed that the intentionality of the agent is irrelevant, what matters is the fact that the outcome is determined by the protected ground (Hacker, 2018, pp. 9-10).

Conversely, indirect discrimination occurs when an apparently neutral provision, criterion or practice creates a disadvantage for persons belonging to a protected group. Nevertheless, a legitimate aim can justify it, considering that the means for achieving it are suitable, proportionate and necessary.[12] This constitutes the most common scenario when dealing with discriminatory AI. Although the values granted to variables or labels usually obey to neutral rules, discrimination might still take place. In addition, certain ML techniques (such as deep learning) do not allow the developers to recognize the relevant variables that explained a discriminatory decision.

It should be noted also that, even if the protected grounds (such as gender or race) are removed from the model in order to infuse it with an appearance of neutrality, proxy discrimination can still happen (if, as in the above-mentioned example, the model relies on another variable to produce a discriminatory outcome). Therefore, unless an AI system has been specially developed in a way that evidences arbitrariness, discriminatory AI outcomes would amount only to indirect discrimination (Hacker, 2018, pp. 10-12).

This is particularly problematic because indirect discrimination is allowed when certain conditions are fulfilled. (I) Legitimate aim, (II) suitability and (III) necessity are enshrined in the Directives dealing with discrimination, while in the case law of the Court of Justice references are made to (IV) the proportionality of the means.[13]

The requirements of legitimate aim and necessity could hardly constitute a barrier for discriminatory AI. On the one hand, an aim such as measuring the credit worthiness of a client or predicting a convict' risk of reoffending falls under the notion of legitimate aim. On the other, necessity is fulfilled as long as there are no less-discriminating alternatives of achieving the same level of accuracy (Hacker, 2018, pp. 17-18).

Although something similar occurs with suitability, some notes need to be highlighted. In *Seymour-Smith*, the Court argued that "mere generalizations concerning the capacity of a specific measure (…) are not enough".[14] As Hacker highlighted, statistics are inherent to AI systems, where the system's accuracy is constantly measured. To put it another way, it'd be relatively easy to provide concrete measured evidence to the Court concerning the suitability of the system. Therefore, discriminatory AI might be justified on the basis of the model's accuracy. Two possibilities need further consideration.

- If the training data is biased, the accuracy of the system will likely be compromised. At the same time, although a biased training data set would have lower

[11] Art. 2(a). Council Directive 2000/78/EC of 27 November 2000.

[12] Arts. 2(b) & 2(b)(I). Council Directive 2000/78/EC of 27 November 2000.

[13] Case 170/84 of 13 May 1986. *Bilka - Kaufhaus GmbH v Karin Weber von Hartz*, para. 35.

[14] Case C-167/97 of 9 February 1999, para. 76.

levels of accuracy in the real world, that would be hard to demonstrate. Conversely, a biased AI system would still produce accurate results within the biased data set. While it'd be hard to demonstrate for those who have suffered the consequences of the lack of accuracy in the real world, it would be easy to show the model's accuracy within a biased data set.

- In contexts of proxy discrimination, the accuracy of the system is not usually compromised. As it was highlighted above, even if the data sets are correctly developed and the model is tested with successful results, proxy discrimination might still happen in the real world. Although bias might be reduced to the detriment of accuracy, it generally implies technical difficulties and economical costs. In cases of proxy discrimination, it can be established that high levels of the model's accuracy would fulfil the requirement (Hacker, 2018, pp. 18-19).

The element of "proportionality of the means" raises potential questions that remain unanswered. In, *Bilka-Kaufhaus*, the Court determined that "if (…) the measures (…) correspond to a real need (…), are appropriate with a view to achieving the objectives pursued and are necessary to that end,"[15] the fact that the measures affected a greater number of women than men did not amount to an infringement. Following this reasoning, to the extent that an AI system is highly accurate, cases of proxy discrimination would arguably pass the proportionality test. Conversely, in cases where the origin of the discriminatory results lies in the biased training data set, different issues arise. First, it should be noted that an unbiased training data set does not preclude the latter occurrence of proxy discrimination. Second, it shall be considered whether the generated model is accurate and obeys a real need. And third, other circumstantial factors such as the existence (or not) of alternative data sets or the economic cost attached to the reduction of biases should be examined. In sum, depending on the circumstances, even a wrongful handling of training data set might qualify as proportional (Hacker, 2018, p. 19).

One last item should be briefly analyzed in this section: the capacity of an AI system to create new categories of discrimination that do not follow classical typologies. This concept has been addressed as affinity profiling. Wachter defines it as "grouping people according to their assumed interests rather than solely their personal traits" (Wachter, 2020, p. 1). It is often used for targeted advertising purposes, and it collides with several human rights (such as the right to privacy).

It also holds several implications for the EU approach towards discrimination. The grounds protected by EU law were defined by historical reasons. However, the ability of modern AI systems to infer characteristics of a certain group can lead to the creation of *ad hoc* groups that escape classical typologies. These artificially created groups (such as "dog owners" or "people born in March") will suffer discrimination in similar ways to protected groups (Wachter, 2020, pp. 55-56). This is problematic because, as it was established in *Chacón Navas*, the listed discrimination grounds cannot be extended by analogy.[16]

Connected to this, the concept of discrimination by association must be introduced. It entered the EU realm in *Coleman*. The Court recognized here that whenever a person who does not belong to a protected group is discriminated due to an association with it, the act would anyways qualify as direct discrimination.[17] This raises several questions regarding profiling techniques. Arguably, if a user who is "profiled" into a protected group – given his or her assumed interests – is discriminated, that person could be entitled to bring a claim (Wachter, 2020, pp. 8-9).

All these implications arising from the EU approach towards non-discrimination must always be examined in the light of trade secrecy law implications. The scarce case-law shows that full disclosure of an algorithm almost constitutes a *rara avis*, greatly hampering the protection granted by anti-discrimination EU law.

[15] Case 170/84 of 13 May 1986, para. 36.
[16] Case C-13/05 of 11 July 2006, para. 56.
[17] Case C-303/06 of 17 July 2008, para. 56.

2.3. Doing AI the European way. The Commission' proposal

The link between the Commission' proposal and the defence of fundamental rights is undeniable – in fact, the term appears 80 times along the document. However, is it enough to address the legal gaps highlighted above?

The Commission defines its approach as risk-based. Accordingly, the proposal differentiates between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk. Uses falling under the first category were directly forbidden. The second one involves the fulfilment of a set of legal requirements, whereas the last one only entails certain transparency obligations.[18]

The proposed AI Act Annex III contains a list of AI uses classified by areas (that vary from "law enforcement" to "access and enjoyment of essential private services") that would qualify as high-risk. This is of great relevance for the purposes of this research, considering the discriminatory potential of many uses listed there. Among others, the Annex refers to systems employed for measuring credit worthiness of natural persons as well of those "intended to be used for making decisions on promotion and termination of work-related contractual relationships."[19] Art. 7 establishes the Commission's faculty of adding new entries into the list. This approach was also followed by the UNESCO Ad Hoc expert group, that, attending to the changing nature of AI, decided to focus their work on the relevant features of AI instead of the notion itself (Ad Hoc Expert Group for the preparation of a draft text of a recommendation on the ethics of artificial intelligence, 2020, p. 4). It has the advantage of allowing the Commission to react relatively fast to the challenges introduced by innovation.

High-risk AI systems would have to meet a set of requirements established in the proposal Art. 10. Art. 10(3) establishes that training data sets "shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used." This provision is complemented by Art. 10(4), where the same data sets are required to "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 Commission puts the focus not only on the representativity of the data but also on (i) the groups of people targeted by the system and (ii) the environmental circumstances surrounding the display of the system.

Nevertheless, although the generic wording seems to obey the (increasing) wide variety of AI uses, a broad obligation of representativity leaves the door open for more questions. On the one hand, because no threshold of representativity is established in the proposal. On the other hand, because in many cases representative data will be inexistent. Or its development might imply a higher economical investment for the company, potentially hindering innovation.

What about proxy discrimination? Well, that is somewhat addressed *ex-post*. Art. 12 establishes a design-obligation for AI systems. They must be built with the capacity to record incidents "with respect to the occurrence of situations that may result in the AI system presenting a risk."

Another safeguard contained in Art. 14 complements this provision. It is established that AI systems must be designed "in such a way (…) that they can be effectively overseen by natural persons." The human supervision must be designed in a way that allows the fulfilment of a set of functions contained in Art. 14(4). Among them, it must allow the supervisor to "remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system' and to 'be able to (…) disregard, override or reverse the output of the high-risk AI system."

Although the highlighted measures will not fully prevent the occurrence of discriminatory outputs and proxy discrimination, they could have the potential to facilitate its

[18] Arts. 5-6. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act), COM/2021/206 final.

[19] Annex III. Artificial Intelligence Act.

detection and (at least) diminish its negative consequences. It is important to emphasize the "could" because these obligations are intended to be enforced through a self-assessment procedure[20] called conformity assessment.[21] As the NGO Algorithm Watch highlighted (Reinhold, 2021), corporate actors will be interested on deploying the systems into the market. Therefore, the question of whether an interested part can assess objectively the fulfilment of the highlighted obligations remains.

Concerning the opacity of the systems, the proposal tries to increase transparency while respecting IP rights. On the one hand, the proposal advocates for the creation of an EU Database for stand-alone High Risk AI systems. However, the information that must be provided does not cover the functioning of the system and has been contested. Algorithm Watch criticized that is lacking "an explanation of the model (logic involved) and details on who developed the system, as well as the results of any algorithmic impact assessment/human rights impact assessment undertaken by public authorities" (Reinhold, 2021). *A priori*, it seems that, although certain transparency obligations have been established, the barrier that IP rights constitute will remain.

Conclusion

It cannot be stated without hesitation that the established EU human rights legal framework is able to cope with the challenges for non-discrimination introduced by AI. To the date, there is no foolproof methodology that can face the highlighted challenges introduced by AI. On the one hand, because the existence of IP rights greatly hinders the acknowledgement of an AI discriminatory output. On the other hand, because, in the proposed AI Act, the possibility of overcoming challenges depends to a large degree on whether the developing companies are committed to fulfil the requirements established in the proposal.

However, it is also somewhat clear that the functioning of the EU human rights legal framework leaves the door open for further developments. And this is a fundamental advantage. In traditional IHRL, the debate is still focused on whether non-state actors can be found accountable for human rights violations (Suppa & Bureš, 2020, pp. 153-179). Within the EU realm, the debate has gone way beyond that "starting" point. EU law arguably holds a more dynamic character than IHRL, and dynamism is fundamental when dealing with disruptive technologies.

Therefore, even if nowadays the established EU framework is not fully able to cope with the challenges, it is still able to evolve in order to address them. This is where the added value of the EU human rights legal framework lies.

[20] See, at this regard, IOANNIDIS, N.; GKOTSOPOULOU, O. (2021, July). "The Palimpsest of Conformity Assessment in the Proposed Artificial Intelligence Act: A Critical Exploration of Related Terminology" [online]. Available at: https://europeanlawblog.eu/2021/07/02/the-palimpsest-of-conformity-assessment-in-the-proposed-artificial-intelligence-act-a-critical-exploration-of-related-terminology/ [Accessed: 13 August 2021].

[21] Art. 3 (20). EUROPEAN COMMISSION (2021), supra note 56.

References

Scientific literature

ANDERSEN, L. (2018). "Human Rights in the age of Artificial Intelligence". In: *Access Now*, pp. 1-40 [online]. DOI: https://doi.org/10.1007/978-1-4842-3808-0_1

AIZENBERG, E.; VAN DEN HOVEN, J. (2020). "Designing for human rights in AI". In: *Big Data & Society*, pp. 1-14 [online]. DOI: https://doi.org/10.1177/2053951720949566

GRUODYTĖ, E.; MILČIUVIENĖ, S.; PALIONIENĖ, N. (2020). "The Principle of Direct Effect in Criminal Law: Theory and Practice". In: *European Studies*, vol. 7, pp. 66-87.

HACKER, P. (2018). "Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU Law". In: *Common Market Law Review*, vol. 55, no. 4, pp. 1143-1183.

HUSEINZADE, N. (2021). "Algorithm Transparency: How to eat the cake and have it too". In: *European Law Blog* [online]. Available at: https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/ [Accessed: 15 April 2021].

IOANNIDIS, N.; GKOTSOPOULOU, O. (2021). "The Palimpsest of Conformity Assessment in the Proposed Artificial Intelligence Act: A Critical Exploration of Related Terminology". In: *European Law Blog* [online]. Available at: https://europeanlawblog.eu/2021/07/02/the-palimpsest-of-conformity-assessment-in-the-proposed-artificial-intelligence-act-a-critical-exploration-of-related-terminology/ [Accessed: 13 August 2021].

KOCHLING, A.; WEHNER, M. C. (2020). "Discriminated by an algorithm: a systemic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development". In: *Business Research*, vol. 13, pp. 795-848 [online]. DOI: https://doi.org/10.1007/s40685-020-00134-w

LIU, H.; DANAHER, J. et. al. (2020). "Artificial Intelligence and Legal Disruption: A New Model for Analysis". In: *Law Innovation and Technology*, vol. 12, no. 2, pp. 1-46.

RASO, F. *et al*. (2018). "Artificial Intelligence & Human Rights: Opportunities & Risks". In: *Berkman Klein Center for Internet & Society at Harvard University*, pp. 1-63 [online]. DOI: https://doi.org/10.2139/ssrn.3259344

SUPPA, A.; BUREŠ, P. (2020). "Can Multinational Corporations be responsible for human rights violation of its outsource company? Response of national or international law?". In: *International and Comparative Law Review*, vol. 20, no. 1, pp. 153-179 [online]. DOI: https://doi.org/10.2478/iclr-2020-0007

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. (2017). "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". In: *International Data Privacy Law*, vol. 7, no. 2, pp. 76-99 [online]. DOI: https://doi.org/10.1093/idpl/ipx005

WACHTER, S. (2020). "Affinity Profiling and Discrimination by Association in Online Behavioural Advertising". In: *Berkeley Technology Law Journal*, vol. 35, no. 2, pp. 1-73.

Books

HAMULAK, O.; STEHLÍK, V. (2013). *European Union Constitutional Law: Revealing the Complex Constitutional System of the European Union*. Olomouc: Palacký University Olomouc.

RAMIRO TROITIÑO, D.; KERIKMÄE, T.; CHOCHIA, A. (2020). "Foreign Affairs of the European Union: How to Become an Independent and Dominant Power in the International Arena". In: *The EU in the 21st Century*, pp. 209-230 [online]. DOI: https://doi.org/10.1007/978-3-030-38399-2_12

RAMIRO TROITIÑO, D.; KERIKMÄE, T.; DE LA GUARDIA, R.; PÉREZ SÁNCHEZ, G. (eds). *The EU in the 21st Century* [online]. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-38399-2_12

Legal sources

Legal norms

C:2008:115:TOC. *Consolidated Version of the Treaty on European Union*.

EUR-Lex-52021PC0206. *Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*.

EUR-Lex-52021PC0206. *Annexes to the Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*.

EUR-Lex-32000L0078. *Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation*.

EUR-Lex-32016L0943. *Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure*.

2000/C 364/01. *Charter of fundamental rights of the European Union. Luxembourg, Office for Official Publications of the European Communities*.

EUR-Lex-52008PC0426. *Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation COM/2008/0426 final*.

EPC1973. *Convention on the Grant of European Patents of 5 October 1973*.

EUR-Lex-32016R0679. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*.

Case-law

Case 43/75 of 8 April 1976. *Defrenne v. Sabena* (no. 2).

Case 170/84 of 13 May 1986. *Bilka-Kaufhaus GmbH v Karin Weber von Hartz*.

Case C-167/97 of 9 February 1999. *Regina v Secretary of State for Employment, ex parte Nicole Seymour-Smith and Laura Perez*.

Case C-13/05 of 11 July 2006. *Sonia Chacón Navas v Eurest Colectividades SA*.

Case T-201/04 of 17 September 2007. *Microsoft Corp. v Commission of the European Communities*.

Case C-303/06 of 17 July 2008. *S. Coleman v Attridge Law and Steve Law*.

Other sources

UNESCO. *Artificial Intelligence* [online]. Ad Hoc Expert Group for the preparation of a draft text of a recommendation on the ethics of artificial intelligence. First Draft of the Recommendation on the

Ethics of Artificial Intelligence in UNESCO, Paris, 7 September 2020. Available at: https://en.unesco.org/artificial-intelligence/ethics#aheg

AMAZON. *What is data labeling for machine learning?* [online] . Available at: https://aws.amazon.com/es/sagemaker/groundtruth/what-is-data-labeling/ [Accessed: 13 August 2021].

European Commission White Paper on Artificial Intelligence: a European approach to excellence and trust [online] [2020]. Available at: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Patent Office, Guidelines for Examination, 3.3.1 Artificial Intelligence and Machine Learning [online] [Accessed: 15 April 2021]. Available at: https://www.epo.org/law-practice/legal-texts/html/guidelines/e/g_ii_3_3_1.htm

DASTIN, J. (2018). "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* [online]. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Accessed: 20 April 2021].

DILMEGANI, C. (2021). "Data Labeling in 2021: How to Choose a Data Labeling Partner". In: *AI Multiple* [online]. Available at: https://research.aimultiple.com/data-labeling/ [Accessed:13 August 2021].

Furtwangen University. Machine Learning Work Group [online]. Available at: https://www.hs-furtwangen.de/en/faculties/computer-science/research/translate-to-english-machine-learning/ [Accessed: 11 September 2021].

HENLEY, J.; BOOTH, R. (2020). "Welfare surveillance system violates human rights, Dutch court rules". In: *The Guardian* [online]. Available at: https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules [Accessed: 20 April 2021].

LISA COLE, N. (2019). "Understanding segregation today". In: *ThoughtCo* [online]. Available at: https://www.thoughtco.com/understanding-segregation-3026080 [Accessed: 14 August 2021].

REINHOLD, F. (2021). "Algorithm Watch's response to the European Commission's proposed regulation on Artificial Intelligence – A major step with major gaps". In: *Algorithm Watch* [online]. Available at: https://algorithmwatch.org/en/response-to-eu-ai-regulation-proposal-2021/ [Accessed: 1 May 2021].

XORIANT. *Decision Trees for Classification: A Machine Learning Algorithm* [blog]. Available at: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html [Accessed: 13 August 2021].

ZHANG, M. (2015). "Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software". In: *Forbes* [online]. Available at: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=26a7b5f4713d [Accessed: 10 April 2021].

Recommended citation

MARTÍNEZ-RAMIL, Pablo (2021). "Is the EU human rights legal framework able to cope with discriminatory AI?". *IDP. Internet, Law and Politics E-Journal*. No. 34. UOC [Accessed: dd/mm/aa]
http://dx.doi.org/10.7238/idp.v0i34.387481

About the authors

Pablo Martínez-Ramil
Department of International and European Law. Faculty of Law, Palacký University Olomouc
pablo.martinezramil01@upol.cz

Ph.D. Researcher at the Department of International and European Law. Faculty of Law, Palacký University Olomouc. He holds a Bachelor's Degree in Political Science and Public Administration from the University of Salamanca (2012-2017), a Bachelor's Degree in Law from the University of Salamanca (2012-2017). He also holds a Master's Degree in Political and Social Leadership from the University Carlos III of Madrid. (2017-2018) a Master's Degree on International and European Law by the Palacky University of Olomouc (2018-2020) and finally a Ph.D. on International and European Law by the Palacky University of Olomouc (2020-2021). Pablo has previous experience as a Public Affairs Consultant in Lasker Integrated services. Lobbying and Institutional Relations (2018). He is also a social activist and a volunteer: International Amnesty in the regional structure of Castilla y León (2018), Legal Clinic of Social Action. University of Salamanca (2017), Curricular practices in the local group of International Amnesty in Salamanca (2016) and Voluntary Service in the Red Cross (2015). Over the years he has won the following awards: Master on Political and Social Leadership' Extraordinary Prize granted by the University Carlos III of Madrid. 2018 and Member of the Czech team in the 2020 Telders International Law Moot Court Competition.