Journal of Ethics and Legal Technologies – Volume 4(1) – May 2022

Discriminatory algorithms. A proportionate means of achieving a legitimate aim?

*Pablo Martínez-Ramil**

Palacký University Olomouc
pablomramil@usal.es

Abstract: Within the EU legal framework, cases of indirect discrimination are justifiable when suitable and proportionate means are necessary to achieve a legitimate aim. However, the technical reality attached to machine learning (ML) environments challenges this legal postulate from multiple angles. Previous academic writings suggest that, in most scenarios, discriminatory outputs produced by an AI system would meet not only the requirements for indirect discrimination but also the requirements for its justification. This research confirms those views by analysing the relevant case law of the Court of Justice of the European Union (CJEU) to determine the current standing of the prohibition of indirect discrimination. The analysis is complemented by a case study in which it is discussed whether scenarios of proxy discrimination caused by an automatized recruitment tool would pass the CJEU test.

Keywords: indirect discrimination, Machine Learning, Artificial intelligence, proxy discrimination

* Pablo Martínez Ramil worked on this paper on behalf of the Project of specific research no. IGA_PF_2021_013 "Ensuring Human Rights' protection in the European Union in the Age of Artificial Intelligence

1. Introduction

The anti-discrimination legal framework applicable in Europe is far from simple. From a holistic perspective, it is composed of the intersection of different legal regimes: the European Union (hereinafter EU) anti-discrimination law, the European Convention on Human Rights, the relevant provisions within the national legal systems, and the obligations emanating from International Law (European Union Agency for Fundamental Rights, 2018, pp. 16-17 & 24-26).

Despite the persistent debate (see, in this regard, Forshaw & Pilgerstorfer, 2008, and Yu, 2019), most legal systems acknowledge the distinction between direct and indirect discrimination. While the former refers to conducts, policies or practices that treat people differently based on a protected ground (such as race, age or sexual orientation), the latter encompasses practices that, imbued with an appearance of neutrality, put persons sharing a specific protected feature at a particular disadvantage (Lane & Ingleby, 2017, pp. 531-532). This distinction is of great relevance, for conducts or practices qualifying as indirect discrimination might find justification as long as the practice constitutes a proportionate mean of achieving a legitimate aim.

Hence, as long as a valid objective is pursued, the high accuracy of some types of Artificial Intelligence (hereinafter AI) systems with discriminatory potential would allow the system to escape the prohibition. Especially, if the alternative (adjusting the inner parameters to develop a less discriminatory algorithm) involves higher costs and diminishes the system's accuracy (Hacker, 2018, pp. 16-22).

Although previous investigations have already explored this issue (see, in this regard, Hacker, 2018; Martínez-Ramil, 2021), this research aims to go one step further and take a closer look at the cornerstones of this proportionality test, considering not only the current understanding of the notion of indirect discrimination but also the latest legislative developments. To do that, this article will first introduce some relevant features of the technology and the applicable law. The analysis will continue examining the relevant case law of the CJEU to present a contemporary understanding of the concept of indirect discrimination. The examination will conclude with the fourth section, which will explore the issues that arise when a case of indirect discrimination has its origins within an algorithm, using a case study of a discriminatory automated recruitment system (hereinafter ARS) to illustrate them.

2. Preliminary notes on AI and anti-discrimination law

The range of questions that AI possess for anti-discrimination law grows by the day, in line with the increasing pace of technological innovation (Liu et al, 2020, pp. 205-206). A comprehensive study of the emerging issues would justify more than an article. Therefore, the scope of this research has been narrowed to address in detail a specific question: In which scenarios discriminatory ML systems can circumvent the prohibition of indirect discrimination enshrined in the EU anti-discrimination directives?

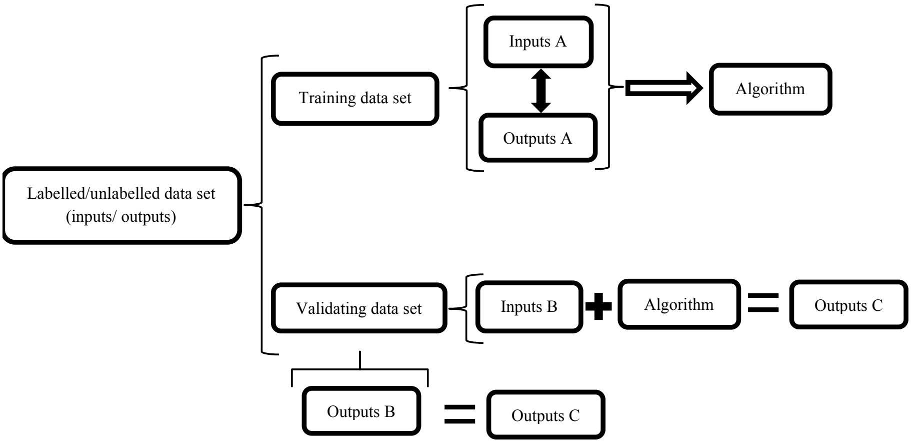
2.1. A brief roadmap on AI systems

Although the present section will provide the reader with some contextual information on the topic, it is worth recalling that previous investigations have thoroughly addressed the functioning of AI systems and their discriminatory potential (see, in this regard, Hacker, 2018; Wachter, 2020).

For the purposes of this study, the definition enacted in the proposed *AI Act* will be taken into consideration. The European Commission defined AI systems as "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs (...)." The aforementioned Annex classifies the AI developing techniques into ML approaches, logic and knowledge-based approaches, and statistical approaches. In simple terms, ML function as an umbrella term that encompasses those algorithmic models that allow the AI to learn "by example" (Hacker, 2018, p. 5). Its widespread use and popularity (Sarker, 2021, pp. 1-2) validate this article's choice of examining their features in light of the CJEU understanding of indirect discrimination. To that end, it becomes necessary to grasp some of the science behind ML environments.

In a few words, a ML system is software that interprets data, using a mathematical formula (so-called "algorithm") to produce a result. The procedure by which a ML system is typically developed appears elucidated in *Figure 1.* In an early stage, the data is divided into two sets: training data and validating data. The first is composed of both inputs (A) and outputs (A). Using an ARS as an example, the inputs would include all the data obtained from job applications, while the outputs would be composed of the recruitment decisions (whether the applicants were hired or not). At that point, the system processes the data and creates an algorithm that explains the relationship between inputs and outputs. The algorithm is then tested through the Validating data set. This time, only the inputs (B) are introduced and it is let to the algorithm to elaborate the outputs (C). Revisiting the ARS example, the system will now freely decide whether to hire an applicant or not. If the outputs (C) equal the outputs (B) (i.e., if the algorithm successfully predicts whether the applicants were hired or not), the system is considered ready for its deployment (Hacker, 2018, pp. 5-6).

Figure 1



A basic representation of the development of an ML environment.

Two typologies of ML systems are of great relevance to this text.

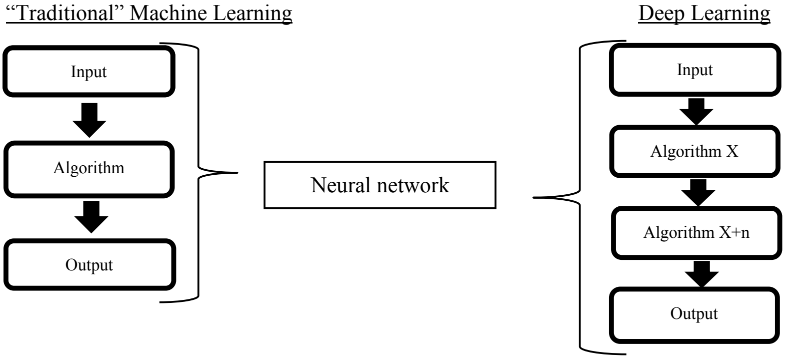
Depending on whether the data is labelled (if informative labels have been assigned to the raw data to influence the interpretation capacity of the system), three main types[1] are recognized within the introduced scheme: supervised (if the data is labelled), unsupervised (if the data isn't labelled) and semi-supervised (if the data sets contain both unlabeled and labelled data) (Zeng & Long, 2022, p.4).

Typically, a ML environment comprehends three layers: the input (request), the algorithm (processing) and the output (result). This is referred to

[1] Most classifications include "reinforcement learning." Its absence here obeys to (i) its different functioning and (ii) the current applications of this technology. Conversely to the scheme drafted in Figure 1, reinforcement learning systems develop strategies for solving problems interacting with the einvornment. The training data is obtained through these interactions, that seek to obtain a reward from the environment (Zeng & Long, 2022, p. 601). It is possible to illustrate this in easier terms with a not very practical example: a system employed for advertisement purposes that displays a banner on a website. Every click obtained would constitute a reward for the system, and the strategies developed to achieve this aim (the organization of the words and pictures within the banner) would constitute the training data. A system like the one described here would need large amounts of training time (a classical feature of reinforcement learning) while being exposed to fake negatives that could negatively impact the evolving algorithm (like misclicks). In other words, a discriminatory output would not be explained by similar channels like the ones discussed above. The current applications involve the gaming industry, healthcare, self-autonomous vehicles… (Mwiti, 2021).

as a neural network. However, when dealing with large amounts of data sets, some systems dedicate several layers to the data processing phase (Anrig et al, 2008, p. 77). The addition of layers radically increases the complexity of the processing and, although this improves the performance of the system, it also hinders the traceability of a certain output (in other words, the reasons that lead the AI system into a determined result) (Hoepman, 2018, p. 50). This is known as Deep Learning (hereinafter DL). DL can also be supervised, semi-supervised or unsupervised.

Figure 2



Simplification of the neural networks of traditional ML and Deep Learning.

All types of ML environments are at risk of producing discriminatory outputs if the data sets used in the ML development procedure are flawed (i.e. if it is incomplete, contains errors, irrelevant information, statistical bias...) (Andersen, 2018, pp. 9-12). Moreover, supervised and semisupervised ML might contain errors related to the labelling of the data. And just as importantly, in unsupervised and DL environments, the system might decide on the outcome of a request based on data directly or indirectly linked to a ground of discrimination (Xenidis, 2021, p. 746). The latter case is referred to as "proxy discrimination." The particular features of this process depend on the complexity of the system (i.e. number of layers of the neural network); and deserve a closer examination.

2.2. The cornerstones of proxy discrimination

The increasing relevance of the concept of proxy discrimination in ML environments provoked some academic reactions on both sides of the Atlantic. In Europe, Hacker (2018, p. 6-7) defined it as the situation "in which precise information about the desired trait (...) is lacking and the decision-maker therefore substitutes the desired parameter with an easily observable one." Likewise, Xenidis (2020, p. 746) considers it as "discrimination based on correlation with protected grounds." Both definitions derive from tackling the core elements of the concept in analogue environments. In the United States (hereinafter the US), Schwarcz (2021, p. 106) established that it is only possible to talk about proxy discrimination when two conditions are met. First, apparently neutral information must disproportionally harm members of a protected group. Second, the statistical value of the apparently neutral information is significantly grounded on its capacity to proxy for a protected ground.

Under the EU approach towards indirect discrimination, what matters is the neutrality of the practice and the disparate impact on a protected ground (an issue that will be explored below). The second part of the definition proposed by Schwarcz, although fundamental to the study of the theoretical distinction (E.R. Prince & Schwarcz, 2020, pp. 1260-12602), becomes irrelevant to the purposes of this research. Hence, this article will understand (in ML contexts) proxy discrimination as the process by which a ML system disproportionately generates negative outputs for members of a protected group based on apparently neutral data correlations.

The notion is not as straightforward as it might seem at first glance. A textbook example could involve an algorithm relying on the variable "postal code" to produce discriminatory outputs. Acknowledging a potential correlation between an ethnic group and the land where it rests, to a certain extent a link can be presumed between the protected group and the postal code they share. Hence, a likely neutral variable ("postal code") determines a differentiated impact against a protected ground ("ethnic minority"). In other words, the variable "postal code" becomes a proxy for "ethnic."

In reality, crystal clear cases of proxy discrimination like the one exposed here would rarely escape the human eye. However, its simplicity evidences the mechanism behind the process. In the context of an ARS, the historical discrimination suffered by ethnic or racial minorities has its reflection on their employment rates. Thus, even if a variable registering a protected value is removed from the data sets (either in compliance with a legal obligation or by decision of the AI developer), the fact remains that the protected ground (call it "race" or "ethnic") was the variable that explained why some people were hired and some not. So, in absence of the explanatory variable, the system will use the training data to identify by itself proxies for that missing ground (in this example, "postal code").

As it was observed here, the chances of proxy discrimination occurring increase when the protected ground effectively predicts the relationship between inputs and outputs within the training data set. Specifically, if there is no more accurate alternative data available (E.R. Prince & Schwarcz, 2020, pp.1263-1265).

Prince & Schwarcz (2020, pp. 1277-1281) distinguished three types of proxy discrimination. The first one was labelled as "casual proxy discrimination" (hereinafter CPD) and it would mirror the example presented above. One apparently neutral variable ("postal code") proxies for a protected ground ("ethnic") to produce the outcome ("not to hire").

The second type was named "opaque proxy discrimination;" (hereinafter OPD) and it operates slightly differently. In the previous case, the protected ground was the causal explanation of the outcome. In absence of it, the ML system uses an apparently neutral variable to proxy it. In OPD scenarios, the protected ground is correlated with the outcome; but it is not its cause. The correlation emanates from unquantifiable or unavailable data; and the protected ground functions as a proxy for them. A case involving both unavailable and unquantifiable data would address gender disparities in mortality. Some of the reasons that explain why women tend to live longer than men are unquantifiable (environmental and biological factors) or unavailable (if, for the purposes of this example, it is assumed that other unknown reasons apply). Hence, to predict mortality, a ML system might rely on an apparently neutral variable ("shopping history") to proxy for a protected ground ("sex") that proxies for other unavailable and unquantifiable data.

The third category is entitled "indirect proxy discrimination" (hereinafter IPD). In IPD cases, a protected ground becomes predictive of a certain output because it proxies for an external quantifiable and available variable causing the outcome. These instances are generated due to the existence of incomplete data (either in the inputs or in the training data set). It can be illustrated in the context of an ARS. According to correlations identified in the training data set, height (explanatory variable) is predictive of better job performance (desired output). However, the data in the CVs of the job applicants do not contain such information (incomplete data). To overcome this, the system identifies an imperfect correlation between height (explanatory variable) and sex (protected ground). Thus, a proxy for sex (such as "Netflix viewing habits") would become predictive for "better job performance." Scenarios of IPD will not take place as long as the ML system has access to (i) the missing information (height) or (ii) better proxies (for instance, if the system has access to "clothes shopping history").

The materialization of the three types of proxy discrimination introduced here is strongly influenced by the type of ML system that produces them. Therefore, models functioning with smaller amounts of data (e.g., certain supervised ML techniques, like decision trees) would generally not (theoretically) present strong complications regarding the identification of cases of proxy discrimination. Conversely, the ones working with large amounts of data sets (e.g., Deep Learning) require a deeper examination that will be discussed below.

2.3. The scope of the EU anti-discrimination directives

Bearing in mind the potential challenges brought by AI systems, the European Commission highlighted on its White Paper on AI (a document that drafted the policy options for AI and set the grounds for the development of the AI Act), the EU regulative framework remains applicable "irrespective of the involvement of AI" (European Commission, 2020). Considering that previous academic works have already examined the intersection of other instruments of the European anti-discrimination legal framework, the focus of this research will be the concept of indirect discrimination, codified in the anti-discrimination directives.

Nonetheless, before digging into the concept of indirect discrimination, attention should be paid to the applicable levels of protection. As it is exposed in Table 1, a clear distinction arises between the protected grounds of every specific area, being the area of "employment" the one covering a greater range of grounds. The relevance of this distinction lies in the different implications generated for the deployment of AI systems. While an AI system intended to be used as a recruitment tool needs to consider sexual orientation as a protected ground, another one dealing with applications for housing subsidies will not need to.

Table 1

Areas	**Protected ground**
Employment	Race and ethnicity, sex, sexual orientation, disability, religion or belief and age.
Welfare systems	Race and ethnicity
Goods and services	Race and ethnicity, sex
Social security	Race and ethnicity, sex

Scope of application of EU anti-discrimination directives

The notion of indirect discrimination shares a very similar wording in all the anti-discrimination directives. According to them, indirect discrimination "shall be taken to occur where an apparently neutral provision, criterion or practise would put persons" that share a protected feature "at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary."

While keeping that in mind, the small print includes certain exceptions. Different treatment on the grounds of age might be justified if they are "objectively and reasonably justified by a legitimate aim, including legitimate employment policy, labour market and vocational training objectives, and if the means of achieving that aim are appropriate and necessary." In other words, the fulfilment of these conditions precludes the classification of certain conducts as discriminatory on the grounds of age. The directive contains some examples, such as the establishment of a maximum recruitment age (as long as it obeys a legitimate aim, like "the training requirements of the post in question"). The justification for age discrimination constitutes, by its own merits, a topic that would deserve separate research (see, in this regard, Liu, K. & O'Cinneide, C., 2019, pp. 66-67, or De Vos, 2020, pp. 76-29).

The circumstances that justify the conduct of indirect discrimination were drafted in somewhat abstract terminology. Therefore, it is necessary to take a closer look at the judicial interpretation of this prohibition.

3. The current standing of indirect discrimination within the EU realm.

In contexts of direct discrimination, a person is treated less favourably than another on the basis of a protected ground. The very few exceptions to this prohibition have been plainly clarified in the legislative framework. Conversely, the possible justifications for a conduct qualifying as indirect discriminations are almost impossible to list holistically. That is because, if a victim proves discriminatory effects, it is left to the perpetrator to demonstrate that the conduct was proportionate and necessary to achieve a legitimate aim (McCrudden & Prechal, 2009, p. 36). In other words, the acknowledgement of an exception depends on the circumstances of every case.

This section will examine how the CJEU has interpreted the prohibition of indirect discrimination. However, it should be noted that the transposition and further interpretation of the directives have generated different approaches within the EU, where the judicial outcomes strongly depend on the national legal views. This might lead to opposed views on the same conflict depending on, for instance, whether the conflict is being ruled by the United Kingdom's House of Lords -where the so-called *Hampson* test applied-, or by the Court of Justice of the European Union (hereinafter the CJEU) -that implements a stricter threshold- (see, in this regard, Lane & Ingleby, 2017).

Within the EU realm, the first reference to this prohibition can be found in *Bilka*, prior to the elaboration of the anti-discrimination directives. There, the European justice recognized that a practice that "applies independently of a worker's sex but in fact affects more women than men might be regarded as objectively justified economic grounds." The court established for the first time there that it was up to the national court to determine whether "the measure (...) corresponds to a real need on the part of the undertaking, are appropriate with a view to achieving the objectives pursued and are necessary to that end." The elements raised by the Court already resembled in 1986 those of the definition contained in the directives. Here, they will be examined in detail.

3.1. An apparently neutral provision with significantly more negative effects on a protected group.

The case law of the CJEU offers a wide variety of examples of what qualifies as "neutral provision." They could be defined as requirements, conducts, or policies applied to everybody in the same terms. The key that differentiates direct and indirect discrimination lies in the effects. One conduct might qualify as direct discrimination due to its raison d'être (different treatment based on a protected ground). However, indirect discrimination is solely determined by the differential effects, not by the differential treatment (European Union Agency for Fundamental Rights, 2018, pp. 54-56).

Hence, the victim is the one who has to demonstrate that a specific practice puts people sharing a protected feature at a particular disadvantage. The notion of "particular disadvantage" was defined in *CHEZ* (para. 4), in the context of race discrimination (Howard, 2018, p. 64). There, the Court established that the provision "does not refer to serious, obvious or particularly significant cases of inequality, but denotes that it is particularly persons of a given racial or ethnic origin who are at a disadvantage because of the provision, criterion or practice at issue."

The CJEU has recognized the value of statistics for determining this effect along with its jurisprudence. As the Court established in *Seymour-Smith* (para. 57), "the existence of statistically significant evidence is enough to establish disproportionate impact and pass the onus to the author of the allegedly discriminatory measure." The statistics were often used in the case law to compare two groups, being always one at disadvantage and the other one used as a comparator. With this, it should be highlighted that the words' selection "with significantly more negative effects" is not accidental. To be considered indirect discrimination, the measure must produce significant negative effects, but it cannot affect the protected group as a whole. In *Maruko*, a company had refused to pay the complainant the survivor's pension (due to the decease of his partner of the same sex) on the basis of not being married. An action for indirect discrimination was brought before the Court to indirect discrimination since marriage was not an option for same-sex couples in Germany at the time. The Court, using heterosexual couples as a comparator, determined that this practice, despite being neutral in appearance, discriminated against the totality of same-sex couples, amounting to direct discrimination. Hence, in some cases, a neutral provision can also qualify as direct discrimination. This view was supported by later jurisprudence, as in *Frédéric Hay*.

One last case worth mentioning here addresses the concept of indirect discrimination by association. It was acknowledged in the aforementioned *CHEZ*. There, the CJEU determined that the concept of "discrimination on the grounds of ethnic origin" applied also to persons that, even if they did not belong to that group, were nevertheless affected by a discriminatory measure in the same way as someone from the protected category (*CHEZ*, para. 1). Hence, the existence of a link between the discriminatory measure and the racial or ethnic origin should be demonstrated. Depending on the nature of the measure and its effects, a person suffering the measure would be entitled to bring an action for direct or indirect discrimination. In the words of Bruton (2016, p. 14), "this case appears to extend indirect discrimination to the principle of associative discrimination."

3.2. Justification grounds.

Along the case law of the CJEU, it is easy to notice that in many cases it is left to the national court (which is closer to the national reality and the context of every specific case) to determine whether a conduct qualifying as indirect discrimination might find justification. Nonetheless, the repetition of certain rules of interpretation through the case law has originated a certain degree of consensus within academia regarding some concepts.

The first one that will be addressed is the notion of necessity. It is often defined as the absence of less discriminatory alternatives (Hacker, 2018, p. 18). The available case law seems to support this approach. In *Bilka*, the Court established that measure shall constitute a "real need on the part of the undertaking." Thus, it seems that the mere subjective perspective of the perpetrator would not satisfy this condition. Years later, in *CHEZ* (para. 4), the Court established that it "is for the referring court to determine, either that other appropriate and less restrictive means enabling those aims to be achieved exist or, in the absence of such other means, that that measure prejudices excessively the legitimate interest" of the affected population. In other words, the discriminatory measure must constitute a real need and a less discriminatory alternative. Some consider that an obligation emanates from this approach for the respondent, which would have to demonstrate that a less discriminatory alternative would turn inefficient to that end (Euroactiv, 2020, p. 13).

The second is the legitimate aim to which the discriminatory measure obeys. Again, it is not possible to establish a set of rules to interpret this provision, being in most cases left to the national court to determine. In employment matters, it seems that the CJEU has been reluctant to accept defences of employers based on economic concerns, while accepting a differential treatment on cases of positive discrimination (see, in this regard, *María do Ceu* or *Maurice Leone*). However, economic grounds attached to certain positions might be claimed and furtherly admitted if they are properly justified. For instance, a pay practice that, with an appearance of neutrality, disproportionally affects women in a more negative way than men, might be justified by "the state of the employment market, which may lead an employer to increase the pay of a particular job in order to attract candidates" (*Dr Pamela Mary Enderby*, para. 26). Similarly, the Court recognized in *Handels* (paras. 22-24) that specially remunerating "the employee' s adaptability to variable hours and varying places of work [,] (…) the special training [or] (…) the length of service" can be justified -despite qualifying as indirect discrimination- by the specific needs that the performance of a certain position requires.

From a public perspective, the Court established in *Hilde Schönheit* (para. 92) that "restricting public expenditure is not an objective which may be relied on to justify different treatment (…)." Moreover, although the states have a wide margin of appreciation (European Union Agency for Fundamental Rights, 2018, pp. 94-95), in *Ingrid Rinner-Kühn* (para. 14) it was determined that "only generalizations" will not constitute a legitimate aim. It needs to be shown that "the means chosen meet a necessary aim of its social policy and that they are suitable and requisite for attaining that aim."

4. Can discriminatory ML algorithms pass the CJEU test?

This last section will study the suitability of the prohibition of indirect discrimination as interpreted by the CJEU when dealing with potential cases of discriminatory ML systems.

A case study of an ARM will complement the legal analysis. Theoretically, this ARM would be used in recruitment procedures to predict the future job performance of the applicants. This research will consider an unsupervised DL environment (hence identifying patterns and correlations in large data sets through a neural network composed of several layers) engaging in proxy discrimination against people of an ethnic minority.

4.1. Determination of indirect discrimination

An apparently neutral provision, criterion or practice

The appearance of neutrality in ML contexts would generally constitute the rule. However, differences arise depending on whether the approached system is supervised, semisupervised, or unsupervised.

In supervised and semisupervised environments (where a human has introduced labels or assigned a specific weight to a certain feature to influence the functioning of the AI system), the appearance of neutrality might disappear if the variable expressing (not proxying) a protected ground was tampered to hinder it (generating an output that negatively affects the totality of the protected group, bearing in mind the legal justifications of direct discrimination). The resultant algorithm could hardly qualify as neutral (therefore potentially amounting to direct discrimination) (Hacker, 2018, pp. 9-10).

The human role in unsupervised learning environments is significantly smaller. Hence, unless a case involves an algorithm that (i) measures a variable expressing a protected ground in a way that (ii) hinders it and (iii) conditions the discriminatory output, the appearance of neutrality would prevail.

In both scenarios highlighted here, the protected ground, registered in a variable that expresses it within the data sets, justifies the different treatment. In reality, such a case would rarely exist. After all, an AI system that relies on a protected ground to determine the direction of a decision would constitute a terrible predictor. This becomes evident when it is framed in a practical context, as the one suggested here. An ARS system that refuses to hire applicants belonging to a protected ethnic group (hence implying that the protected ground determined the performance of previous applicants) would be de facto disregarding real reasons that explain why some workers were more efficient than others. And, under normal circumstances, the automatisation of the hiring procedures would be motivated by business strategic goals (such as “increasing productivity”), impossible to achieve with an inaccurate system.

Table 2

Type of ML system	**Protected ground**	**Discriminatory outputs**	
		Dependent	**Independent**
Supervised/ Semisupervised ML	Expressed (i.e., "race")	Direct Discrimination (DD)	Indirect Discrimination (ID)
	Proxied (i.e., "postal code")	ID	ID
Unsupervised ML	Expressed	DD	ID
	Proxied	ID	ID

Determination of ID (I). The appearance of neutrality.

The three types of proxy discrimination involve a proxy, a variable correlating with a protected ground. Hence, they maintain an appearance of neutrality (for instance, "shopping history," "postal code," or "member of a Facebook group X") that (at least, at first) could not amount to direct discrimination. An exception might arise in light of *Maruko*, which will be discussed below.

4.2. Significant more negative effects on a protected group

As it was exposed above, the Court has recognised the value of statistics for the identification of disparate impact (*Seymour-Smith*, para. 57). Nevertheless, this is especially problematic in ML contexts due to several reasons. Unless the victim has been able (i) to acknowledge a relevant number of victims of the same protected group or (ii) to obtain an explanation concerning the parameters that determined a significantly negative output, it will be extremely difficult to identify a discriminatory algorithm. The first element is both promising and challenging. It is promising because numbers and statistics are inherent to AI systems. Therefore, an independent auditor would be able to acknowledge a potential problem in compliance with a judicial order or a legislative framework. It is however challenging because, in absence of any legal requirement, the number of victims of the same protected group to identify must be large enough to sustain a claim before a Court, taking into consideration the comparator (e.g. the total number of users of the system). The second element is influenced by technological limits (in deep learning environments, explainable AI constitutes to the day a technical challenge) and the implications of trade secrecy law deserve consideration as well (see, in this regard, Martínez-Ramil, 2021).

Table 3

Identifying disparate impact		**Potential barriers**
Acknowledging a relevant number of victims sharing a protected ground.		In contexts with a large number of users (comparator), this would be rather hard.
Explanation of the parameters governing the algorithm.	Traditional ML	Trade secrecy laws.
	Deep Learning (complex neural networks)	Trade secrecy laws, the opaqueness/ complexity of the correlations.

Determination of ID (II). Disparate impact.

The notion of discrimination by association, acknowledged by the Court in *CHEZ*, gains relevance and complements the above manifested. In ML contexts, any person who shared the feature linked to the discriminatory effects of the algorithm would be entitled to bring a claim. For instance, if that feature were “postal code,” (as in *CHEZ*) anyone sharing the same postal code as the affected ethnic minority would be entitled to initiate judicial procedures.

In addition, the approach established by the Court in *Maruko* and *Frédéric Hay* must not be overlooked in the context of ML environments, for it opens new scenarios to consider. It was established there that conducts with the appearance of neutrality might amount to direct discrimination if it affects the whole of the protected group. The Court determined that members of a protected group (in this case, same-sex couples) were “unable to meet the condition required for obtaining the benefit claimed” (*Frédéric Hay,* para. 44), hence qualifying the provision as direct discrimination. In both cases, it was established that requirements of national law with the appearance of neutrality might amount to direct discrimination if all members of a protected group are unable to meet the condition. A sole application of this approach to the public sphere would disregard the objectives of the EU anti-discrimination legal framework. Hence, assuming its consequent application in private instances, it should be conceded that some manifestations of proxy discrimination resemble the features that decided these cases.

It’s not entirely clear from the wording of the Court whether what is at stake are the discriminatory effects over the totality of the protected group or the material impossibility of the protected group to comply (with the practice, policy or rule). In the first case, all types of proxy discrimination could potentially qualify as direct discrimination as long as the proxying performed by the system prevents the members of a protected group from receiving positive outputs. In other words, when the ML system indirectly discriminates the totality of the group. The second interpretation presents more uncertainties; it involves a prior acknowledgement of the algorithmic elements that hinder the whole protected group. At first glance, this would be impossible in cases of OPD, since the opaqueness of the correlation halts the identification of the unknown or unquantifiable reasons behind the discriminatory outputs. In cases of CPD and IPD, this would depend on the complexity of the system. Systems using simple neural networks and smaller amounts of data would (to a certain extent) allow the traceability of their outputs. Hence, an ex-post analysis of the algorithm mechanism could determine whether a potentially discriminatory output is built upon an unfulfillable condition or just a hindering one for the members of the protected group. Conversely, the complexity of DL environments would hamper the task.

Table 4

Proxy discrimination	***Maruko* and *Frédéric Hay* => Absolute discriminatory effects**	***Maruko* and *Frédéric Hay* => Material impossibility to comply**	
		Traditional ML	**Deep Learning**
Opaque Proxy Discrimination	Possible (Very low risk)	Impossible	
Causal Proxy Discrimination	Possible (low risk)	Possible (ex-post analysis)	Impossible
Indirect Proxy Discrimination	Possible (Very low risk)	Possible (ex-post analysis)	Impossible

Determination of ID (III). When the disparate impact amounts to DD.

Regardless of which interpretation prevails, very rarely an algorithm would establish correlations that negatively affect the whole of a protected group. That is because, when deciding the direction of an output, algorithms do not rely on one specific correlation but rather on a compendium of them, assigning specific weights that vary to every case. Members of a protected group will generally share fewer similarities than differences in their data, which would generally prevent the automatic construction of "absolute" discriminatory models.

As introduced above, the discriminatory ARM (and more specifically, the algorithm that governs it) addressed here would maintain in most scenarios the appearance of neutrality. Despite the difficulties attached to the identification of the discriminatory effects of a measure; it would be very unlikely that the amount of data processed by an unsupervised DL system like the one proposed would generate models absolutely discriminatorily. Especially, considering (i) the large amounts of data processed by DL environments and (ii) the significant role played by the data contained in applicants' CVs in recruitment procedures (for developing an absolute discriminatory model, all the applicants sharing the protected ground should also have very similar information in their CVs).

4.3. Circumventing the prohibition through the proportionality test

Necessity

Necessity was defined in *CHEZ* as the absence of less discriminatory alternatives. In ML contexts, this will likely be the case in most scenarios. As Hacker (2018, p. 18) highlighted, as long as the system has the significant predictive capacity, "its effectiveness will likely surpass any alternative ways of decision making, particularly those based on human decision making unaided by algorithmic computing power."

It should be also noted that measuring the predictive accuracy and the discriminatory potential of an algorithm might be hard to do with data from the real world. In other words, the fact that a system might work perfectly with the training and validating data sets does not preclude the possibility of discriminatory effects when it is deployed for use. Therefore, only certain choices related to a poorly designed algorithm (that is, the selection of an incomplete or biased database to develop the system, as in IPD cases) could be defeated by the requirement of "necessity." In reality, most developers would use complete and unbiased data sets, since it would improve the predictive capacity of the algorithm.

Coming back to the proposed example, the condition of necessity would be satisfied as long as the predictive capacity of the algorithm justifies it. This would be the case if the model does not suffer from design choices that affect its accuracy. If a better model or a better data set was not used due to, for instance, the strong investment that involved, then the Court would have to assess, bearing all the circumstances of the case, whether the measure constitutes a real need to that legitimate aim.

4.4. Legitimate aim.

As was highlighted above, no set of rules determines what would qualify an aim as legitimate. Nevertheless, certain arguments of the analysed case law deserve some comments.

First, it is somewhat clear that the court has accepted differential treatment in cases of positive discrimination, so its mathematical translation into an algorithm[2] should not raise many legal concerns. This however carries attached the possibility of reducing the accuracy of the system (hence, making it prone to other errors). Therefore, an implemententaiton of positive discrimination in AI development should be subjected to a monitoring process.

Second, as it was established above, the Court has rejected pure economic grounds such as restricting expenses as a legitimate aim in analogue cases of indirect discrimination. This could potentially affect ML systems whose discriminatory character was explained by the selection of poor data sets (i.e., data sets containing error or missing information) when better ones were available. In other words, to claim that the selection of a poor data set obeyed to budgetary reasons would not constitute a strong defence. However, as was already mentioned, this will rarely be the case. Developing companies will be interested in developing a high-quality system, and the quality of the data greatly determines the accuracy of the product.

Third, the Court established that mere generalizations do not constitute a legitimate aim. This requirement, in most cases, would not constitute a barrier. There is plenty of scientific evidence that demonstrates the better performance of AI systems in comparison to humans in many different scenarios. Moreover, when an agent is using an AI system, it will have to obey a specific purpose.

Considering the proposed example, as long as the AI developer did not poorly design the AI due to budgetary purposes, this would not constitute a barrier. Moreover, about the third point, the evidence that demonstrates the existence of unconscious bias in recruitment procedures led by humans could support its use, rather than just a generic objective (such as "improving the efficiency of the recruitment procedures").

5. Where do ML systems stand? Conclusions from the case law of the CJEU

In most scenarios, indirect discrimination produced by ML systems will pass the proportionality test of the CJEU. In other words, a discriminatory

[2] For instance, by assigning higher weights to specific values within a decision tree.

algorithm constitutes in most scenarios a necessary means to a legitimate aim. The very few that would not escape the prohibition are (i) those in which the functioning of the algorithm amounts to direct discrimination, (ii) those that involve blatant poor design choices (like the use of biased or incomplete data sets) and (iii) those where the existence of an available less discriminatory alternative prevents the justification.

Moreover, the problem represented by proxy discrimination scenarios will be increasingly gaining relevance in the near future. Whilst is true that the existence of new types of data and the improvement of the already available one will potentially reduce the occurrence of cases, the progressive incorporation of ML into many spheres of everyday life opens the gates for new scenarios.

The legal gaps introduced here are the product of an analogue interpretation of the notion of discrimination. Thus, the digital translation of these concepts will require considerable efforts on the side of the regulator. First, not all challenges can be tackled through legal initiatives. For instance, OPD incidents will depend on the availability and measurability of certain types of data. Second, any legal response must be wide enough to cope with the fast pace of innovation. In this regard, the establishment of data quality standards might help tackle cases of IPD. Likewise, mandatory ex-ante and ex-post analysis can help with the identification of discriminatory correlations in non-complex ML environments. Third, practical solutions to certain issues might not be available in the current state of the art. And, when the solutions are available, new problems might arise. The legislator needs to be aware of all these issues, generated in the digital nature of AI.

Future investigations will deal with the upcoming regulatory instruments and how they deal with the challenges highlighted here.

References

Andersen, L. (2018). Human Rights in the Age of Artificial Intelligence. *Access Now*, pp. 1-40.

Anrig, B., Browne, W. & Gasson, M. (2008). The Role of Algorithms in Profiling. In M. Hildebrant & S. Gutwirth (eds.), *Profiling the European Citizen* (pp. 65-89), Springer, Dordrecht.

Bruton, C. (2016). EU Anti-discrimination law: Definition of key concepts. *Academy of Law (ERA),* Trier, pp. 1-19.

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

Case 171/88. Judgment of the Court (Sixth Chamber) of 13 July 1989. Ingrid Rinner-Kühn v FWW Spezial-Gebäudereinigung GmbH & Co. KG.

Case C-127/92. Judgment of the Court of 27 October 1993. Dr. Pamela Mary Enderby v Frenchay Health Authority and Secretary of State for Health.

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

Cases C-4/02 and C-5/02. Judgment of the Court (Fifth Chamber) of 23 October 2003. Hilde Schönheit v Stadt Frankfurt am Main (C-4/02) and Silvia Becker v Land Hessen (C-5/02).

Case C-267/06. Judgment of the Court (Grand Chamber) of 1 April 2008. Tadao Maruko v Versorgungsanstalt der deutschen Bühnen.

Case C-267/12. Judgment of the Court (Fifth Chamber), 12 December 2013 Frédéric Hay v Crédit agricole mutuel de Charente-Maritime et des Deux-Sèvres.

Case C-173/13. Judgment of the Court (Fourth Chamber), 17 July 2014. Maurice Leone and Blandine Leone v Garde des Sceaux, ministre de la Justice and Caisse nationale de retraite des agents des collectivités locales.

Case C-83/14. Judgment of the Court (Grand Chamber) of 16 July 2015 "CHEZ Razpredelenie Bulgaria" AD v Komisia za zashtita ot diskriminatsia.

Case C-238/15. Judgment of the Court (Second Chamber) of 14 December 2016 Maria do Céu Bragança Linares Verruga and Others v Ministre de l'Enseignement supérieur et de la recherche.

Cases C-804/18 and C-341/19 Judgment of the Court (Grand Chamber) of 15 July 2021 IX v WABE eV and MH Müller Handels GmbH v MJ.

Euractiv (2020). Handbook on the racial equality directive. Indepedent Report September 2020, pp. 1-32.

European Commission (2020). *WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust.* COM(2020) 65 final. Retrieved from: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

De Vos, M. (2020). The European Court of Justice and the march towards substantive equality in European anti-discrimination law. *International Journal of Discrimination and the Law, Vol 20(I),* pp. 62-87.

Diretive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation. *Official Journal L 303*, 2.12.2000, p. 16–22.

Directive 2000/43/EC against discrimination on grounds of race and ethnic origin. *Official Journal L 180*, 19/07/2000 P. 0022 - 0026

Directive 2006/54/EC equal treatment for men and women in matters of employment and occupation (recast). *Official Journal L* 204, 26.7.2006.

Directive 2004/113/EC equal treatment for men and women in the access to and supply of goods and services. *Official Journal L* 373, 21.12.2004, p. 37–43.

E.R. Prince, A. & Schwarcz, D. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review, Volume 105*, pp. 1257-1318.

Forshaw, S. & Pilgerstorfer, M. (2008). Direct and Indirect Discrimination: Is There Something in between? *Industrial Law Journal, Volume 37, Issue 4, December 2008*, pp. 347–364.

Hacker, P. (2018). Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law. *55 Common Market Law Review*, pp. *1143*-1186.

Handbook on European non-discrimination law (2018 edition). (2018). *European Union Agency for Fundamental Rights.*

Hildebrandt, M. (2008). Defining Profiling: A New Type of Knowledge? In M. Hildebrant & S. Gutwirth (eds.), *Profiling the European Citizen* (pp. 17-46), Springer, Dordrecht.

Hoepman, J. (2018). Transparency Is The Perfect Cover-Up (If The Sun Does Not Shine). In E. Bayamlioglu, I. Baraliuc, L. A. Wilhelmina Janssens & M. Hildebrant (eds.), *BEING PROFILED: COGITAS ERGO SUM: 10 Years of Profiling of the European Citizen* (pp. 46-51), Amsterdam: Amsterdam University Press.

Howard, E. (2018). EU anti-discrimination law: Has the CJEU stopped moving forward? *International Journal of Discrimination and the Law, Vol 18(2-3),* pp. 60-81.

Lane, J. A. & Ingleby, R. Indirect Discrimination, Justification and Proportionality: Are UK Claimants at a Disadvantage? *Industrial Law Journal, Volume 47, Issue 4, December 2018*, pp. 531–552.

Liu, K. & O'Cinneide, C. (2019). The ongoing evolution of the case-law of the Court of Justice of the European Union on Directives 2000/43/EC and

2000/78/EC. *European Commission, Directorate-General for Justice and Consumers,* pp. 1-104.

Liu, H., Maas, M. M., Danaher, J., Scarcella, L., Lexer, M. & Van Rompaey, L. (2020). *Artificial Intelligence and Legal Disruption: A New Model for Analysis. Law, Innovation and Technology 12, no. 2*, pp. 205–258.

Martínez-Ramil, P. (2021). Is the EU human rights legal framework able to cope with discriminatory AI? *IDP. Internet, Law and Politics E-Journal, no. 34, UOC*, pp. 1-14.

McCrudden, C. & Prechal, S. (2009). The Concepts of equality and Non-Discrimination in Europe: A practical approach. *European Commission: Directorate-General for Employment, Social Affairs and equal Opportunities*, pp. 1-50.

Mwiti, D. (2021, November 8). 10 Real-Life Applications of Reinforcement Learning. Message posted to http://neptune.ai

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI. 2, 160*, pp. 1-21.

Schwarcz, D. (2021). Health-Based Proxy Discrimination, Artificial Intelligence, and Big Data. *Houston Journal of Health Law and Policy, Volume 21, Issue 1*, pp. 95-154.

Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal, Vol. 35*, pp. 367-430.

Xenidis, R. (2021). Turning EU Equality Law to Algorithmic Discrimination: Three Pathway To Resilience. *Maastricht Journal of European and Comparative Law Volume 27, issue 6,* pp. 736-758.

Yu, A. (2019). Direct Discrimination and Indirect Discrimination: A Distinction with a Difference. *Western Journal of Legal Studies Vol. 9 No. 2*, pp. 1-21.

Zeng, X. & Long, L. (2022). Reinforcement Learning. In: *Beginning Deep Learning with TensorFlow* (pp. 601-674). Apress, Berkeley, CA.

Zeng, X. & Long, L. (2022). Introduction to Artificial Intelligence. In: *Beginning Deep Learning with TensorFlow* (pp. 1-46). Apress, Berkeley, CA.