EQUALITY BETWEEN MEN AND WOMEN IN EU LAW

SEMINAR FOR LAWYERS AND JURISTS
Barcelona, 25 and 26 April 2022

Session on "AI and Gender Equality"

*Biases in the use of artificial intelligence for the management of labour relations: analysis from the perspective of EU anti-discrimination law**

Pilar Rivas Vallejo. Professor of Labour and Social Security Law. University of Barcelona

pilar.rivas.vallejo@ub.edu

Summary:

I. CONTEXTUAL INTRODUCTION: RELEVANT CONCEPTS

1. Artificial Intelligence in law

Artificial intelligence is the *set of scientific methods, theories and techniques whose aim is to reproduce, by a machine, the cognitive abilities of human beings*(*European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment* , of 4 December 2018). For the European Economic and Social Council, it is the "discipline which sets out to use digital technologies to create systems capable of autonomously reproducing human cognitive functions, including in particular grasping data, a form of understanding and adaptation (problem solving, automatic reasoning and learning)" (Opinion on *Artificial intelligence: anticipating its impact on work to ensure a fair transition*, point 2.2).

Although there is not yet any legal definition of artificial intelligence, the proposal for a European Union Regulation on Artificial Intelligence of 21 April 2021[1], incorporates the concept for the purposes of said legislation, assimilating it with “software” (in turn comprising various elements in the form of algorithms, Ebers, 2020: 40). Article 3 (definitions) establishes that "an ‘artificial intelligence system’ (AI system) means 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 such as content, predictions, recommendations, or decisions influencing the environments they interact with".

"Data science" is the computational discipline responsible for processing big data in all phases, beginning with collection by means of various techniques (e.g. data mining and reality mining), and for different purposes, such as prediction[2].

Data processing (big data) is understood as "any operation or set of operations which is performed on personal data, such as collection, storage, recording, alteration, retrieval, disclosure, making available, erasure, destruction or performance of logical and/or arithmetic operations on such data" (Article 2(b) of Convention 108 of the Council of Europe, *for the Protection of Individuals with regard to Automatic Processing of Personal Data*, 1981).

---

*This paper forms part of a version to be published in issue 1 of the year 2022 of e-Revista Internacional de Protección Social, entitled "Sesgos de género en el uso de inteligencia artificial para la gestión de las relaciones laborales: análisis desde el derecho antidiscriminatorio", *e-Revista Internacional de la Protección Social.* Universidad de Sevilla, 2022, vol. 7, issue 1, and of the introductory chapter of the work “Discriminación algorítmica en el ámbito laboral: perspectiva de género e intervención” (Rivas Vallejo, P. dir., 2022) within the context of the research project funded by the Spanish Ministry of Science and Education: “Discriminación algorítmica: género y trabajo”, reference PGC2018-097057-B-I00.

[1] Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative effects {SEC(2021) 167 final} - {SWD(2021) 84 final}-{SWD(2021) 85 final}, at https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF.

[2] Datification allows predictions to be built on the basis of the accumulation of data and their quantification, but also increases the margin of inaccuracy. "To data file a phenomenon is to put it in a quantified format so it can be tabulated and analysed" (Mayer-Schönberger and Cukier, 2013: 37).

2. Machine learning

Artificial intelligence may be viewed from different perspectives, such as an approximation emulating human thought, or as software-supported systems which, by interpreting and learning from big data, simplify or automate certain processes by means of algorithms.

We are interested in the analysed effects of the second mode, which may in turn be distinguished into two models (known as top-down, inspired by behavioural psychology and based on an inductive approach through learning with a sample of examples and construction of models; and bottom-up, based on neuroscience, represented by functional computationalism, which does not use models and which includes deep learning, which learns from prior inferences as to data), on occasion in a hybrid structure (Pujol, 2021: 7), because this is what serves to detect behavioural patterns from which conclusions may be inferred, in large sets of data. In turn, it serves to make highly precise predictions applicable to different commercial and business practices, in short, *machine learning*, or "the set of techniques which learn to find patterns without instructions" ("the branch of artificial intelligence related to the function of learning from experience", Pujol, 2021: 8).

These machine learning models nonetheless suffer from three major drawbacks:

A) First, their calculated design: classifying, ordering, predicting and processing big data by means of an algorithm has a "political" basis, since it could present reality in a particular manner (Bucher, 2018), and is open to manipulation by its designers. Thus, "critical decisions are not reached on the basis of the data themselves, but on the basis of the data analysed by algorithms" (Pasquale, 2015: 21).

B) Second, the data architecture they work with, likewise designed in accordance with certain "intentional" parameters in supervised learning, may also be conditioned by the type and quantity of data employed, since they could be incomplete, biased, or in short, insufficient and unrepresentative (e.g. when working with personal characteristics, with the diversity and sufficient representation of subgroups, individuals, origins, etc.) or poorly labelled (since ultimately this task is performed manually by humans, who also include their own prejudices in the labelling)… or misinterpreted (e.g. *confirmation bias*), giving rise to algorithmic biases.

C) The third problem lies in opacity (according to Burrell –2016–, for three reasons: deliberate corporate or public secrecy, technical ignorance of the functional process, and lastly the inherent design of the model and the misalignment between mathematical optimisation and semantic interpretation of the data), heightened by the fact that to a great extent this is closed source software or a "black box system" (which gives rise to options to resolve this, such as open source design or algorithm audits), the complexity of which is based on the interaction between data and code [Burrell, 2016]. Alongside this, its significance for legal purposes is based on the fact that the conclusions generated by algorithms or AI systems have *powerful legitimacy* (Gillespie, 2016), and furthermore unquestionably influence our decisions, and the decisions that others take about us, encouraging some and ruling out others (Mackenzie, 2017, and Laaksonen, Haapoja, Kinnunen & Nelimarkka, 2020).

In summary, machine learning works with data correlation, but not with "causality" (Spiegelhalter, 2014), because it is much easier to correlate than to detect causes (Mayer-Schönberger & Cukier, 2013: 5), since the autonomy of the algorithm (which allows it to learn for itself and refine how it functions) runs in parallel with its opacity, as it makes it hard to establish the connection between the input data and the results it offers, and hence the origin of the decision or choice provided, in other words *the reason why it believes this to be the best choice.* This may give rise to "false positives" or "false negatives" (Flores, Bechtel & Lowenkamp, 2016), in other words erroneous predictions, which are thus discriminatory, as demonstrated by the use of the COMPAS criminality predictor algorithm in the United States, and these are difficult to detect since an algorithm trained this way ultimately becomes a "black box" (Slavin, 2011). This likewise prevents us from understanding the intermediate process

involved in reaching the conclusions based on the inputs processed, because *access to the algorithm's source code does not truly tell us the source of the decision, if the bias lies not in its design but in the data fed into it.* This element takes on particular importance from the legal perspective, where the basis for decisions is fundamental both in justifying their legality and in determining their validity in the light of anti-discrimination regulations.

II. DISCRIMINATORY BIASES IN THE EMPLOYMENT SPHERE DERIVED FROM THE USE OF AI

1. Digitalisation of old prejudices in the light of rules against discrimination

The lack of transparency, or the opacity, inherent to automated mechanisms increases the typical difficulties in bringing claims against company decisions, which has given rise to growing concern as to this issue in the legal sphere[3], although this is nothing new in the area of data mining, or machine learning and artificial intelligence (Hajian, Bonchi & Castillo, 2016). What is clear is the need to address detection and mitigation of such influences from the perspectives of fields other than computational science (Scantamburlo, 2021: 704), and in particular in legal terms, since their impact gives rise to situations of discrimination which generate prejudices towards people in certain contexts, such as employment. And also because the mathematical models to assist (automated) decision-making could conceal social prejudices which would thus be perpetuated (O'Neil, 2017), exponentially projecting their impact, in particular in the world of work[4].

These concealed biases are on occasion derived from a clearly discriminatory purpose (direct discrimination), but in most cases are simply caused by a lack of consideration of their collateral impact. The hypothetical freedom of the algorithm from contamination is presented as an alternative mechanism to human prejudices or biases. However, if this hypothetical model of objectivity is based on "real" historical data (capturing real discriminatory conduct, the reiteration of which over time detected by the algorithm leaves it to identify this as the right decision), its inherent design no longer corresponds to the intended goal of objectivity, but instead pursues efficiency and productivity. And at the same time, when used to reach employment decisions, this may help to evade the efficacy of a company's equality plan, the measures of which may have been constructed specifically to overcome such prior situations which provide the support for the automated decision-making model, which means that from the perspective of equality between women and men, the progressive replacement of traditional decision-making mechanisms with such automated processes assisted by artificial intelligence (AI) clearly runs counter to the agreed measures to overcome inequalities regarding access to employment, and working conditions[5], unless the equality plan itself provides measures to correct the impact of automated tools, which should be a core element in the design of such plans.

The apparently neutral, objective and uncontaminated nature of automated decision-making mechanisms unquestionably operates to the detriment of the protection of the right to equality, in that the deep learning technique it involves, drawing on the input data used by the algorithms to learn for themselves and infer conclusions which are used to inform decisions, prevents a clear connection from being established between the input data (big data) and the conclusions reached by the mathematical model. The perversity of this functionality lies specifically in the difficult in ascertaining where the bias lies in a decision if it is discriminatory (e.g. in a selection process, because it has decided that one individual has "greater professional value" than another), and in detecting the error so as to challenge the corporate decision.

---

[3] Pioneering contributions have been made by Pasquale and Barocas-Selbst: Pasquale, F. (2019); Pasquale, F. (2020) and Barocas, S. & Selbst, A. D. (2016). I would be in particular recommend the papers by R. Xenidis cited in other notes, and those by Zuiderveen Borgesius, F. (2018).

[4] Rosenblat, A. (2018) & Umoja Noble, S. (2018). Asimismo, Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016).

[5] How this is handled may be referred to Rivas Vallejo, P. (2020) and Rivas Vallejo, P., dir. (2022).

2. Impact of AI in the field of paid employment

The increasing technical complexity of work is a phenomenon which runs in parallel with the evolution of technology itself. Artificial intelligence (AI) represents a further step in this natural evolution, and causes performance at work likewise to adapt to the benefits that it offers, although these may likewise entail the incorporation of a degree of disorder in the dynamic of working relationships, with the emergence of new risks tied to both the execution and management of work.

However, what particularly interests us in this new order is the most disruptive element, which lies not in the tools involved in performing the work themselves, except for digital environments or work with cobots, but in the new way in which the working relationship is managed, and in particular the selection, evaluation, oversight and control of those aspiring to a particular position, or those that have already been hired. In practice, human resource management has to a great extent adjusted to methods which facilitate decision-making, both those with a greater impact, and those corresponding to the everyday dynamic of company management. These are automated models based on AI the design of which is aligned with such typical needs of business activities, focused on certain practical functions of the functionality of mathematical models paired with big data in order to make predictions of recommendations as to the best decision in the context analysed by the model, on the basis of machine learning tools. They offer the advantage of incorporating software which is easily acquired and applied, facilitating the chain of tasks that an algorithm can replace without any investment of time, making their use more popular, and dictating that one would in the short term expect them to be applied *en masse*, rather than being confined solely to big businesses capable of financing their design and production, or where the volume of management processes would invite recourse to tools to simplify this, given their utility in reducing employment costs and controlling flows of labour (Nguyen, 2021: 1), maximising efficacy and productivity.

The proliferation of the use of artificial intelligence and the growth of mass data capture have improved techniques and intensified their increasingly widespread application. Companies now turn to human resource information systems (Talla, Workplace Analytics... as it is specific specifically such back office processes that are the first to be automated [Frank, Roehrig & Pring, 2018: 132]), with the capacity to "acquire, store, handle, analyse, retrieve data and distribute all the information computed in previous steps regarding an organisation's human resources" (Pampouktsi, Avdimiotis, Ìaragoudakis & Avlonitis, 2021).

The application within the context of company activities provides support for all manner of administrative tasks, including labour relations, through the techniques of People Analytics, a sub-science of data mining based on artificial intelligence or talent management and evaluation (Bersin, 2019), serving to feed machine learning systems through which an algorithm or set of algorithms draws conclusions and identifies patterns based on the observation of interaction among (mass) data. Such (historical) patterns allow recommendations to be made accordingly, which entails replicating previous decisions which could be conditioned by social prejudices, as well as turning this bias into a model assisting in future decisions. Clearly, this is a tool which works to the detriment of women in particular, in that the historical patterns that are replicated identify gender roles and prejudices towards work by women, who have historically suffered segregation in employment and worse working conditions (which today's equality plans aim to mitigate through corrective measures, which could in turn see their efficacy undermined by being overlaid with automated decisions which pursue precisely the opposite effect).

The applications which may be found within this context occupy a central role in the selection and hiring of workers, but similarly, systems used to evaluate people, and hence to classify and prioritise them, have the necessary versatility to allow such functions also to be used for other purposes within the dynamic of a working relationship: professional promotion, calculation of pay, continuation at the company or redundancy…

This classification and assessment of people could involve the processing of personal data[6], which facilitates the protection of the individuals involved from the perspective of personal data protection, but often the big data use to feed the system (the algorithm) do not have such status, not in this case because they are anonymised data, but because they are not been processed in order to be used with regard to the data subjects themselves, but instead regarding third parties, which means that any biases that might be derived from the conclusions reached by such an algorithm would apply to new subjects, on the basis of the personal data of the latter. This means that, given this individual impact, it could be relevant to apply Regulation 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* [GDPR]. Although the involvement of automated decisions and their discriminatory impact is very much an incidental aspect of this regulation, the future EU Artificial Intelligence Act channelled by means of the "Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts [COM(2021) 206 final – 2021/106 (COD)]", should provide a response to all the current protection gaps and unknowns that EU anti-discriminatory law covers only partially, while establishing a sufficient legal framework to address gender discrimination (but not multiple or intersectional discrimination). This provision specifically views as high-risk AI systems (Article 6.2 and Annex III, paragraphs 1 and 4) those used for biometric identification and categorisation of natural persons, and those *for employment, worker management and access to self-employment*.

3. Use in selection and recruitment processes

The use of robots or algorithms for selection allows psychogenic techniques to be used in the digital interview phase (known as "screening") to capture all information of relevance for the company, which includes neutralisation of the right to lie in the job interview, reinforcing loyalty to the company and opening up the possibility of engaging in profiling based on physical appearance. This thus excludes individuals from the employment market on the basis of the detection of psychological traits meaning they can potentially be "ruled out" (e.g. predictability of a tendency to depression, likelihood of suffering a mental illness that would be "inconvenient" for the company [Tufekci]; or, among other traits, the trade union trends of the candidates contrary to the corporate culture [Gaster, 2020]).

These situations are not new to the world of employee selection. The novelty lies in the discriminatory potential of new systems used to achieve such purposes, based on two key distinctive features: the opacity and reliability on the part of users, together with the capacity to replicate previous (historical) biases and amplify their impact. Bearing in mind that they are used before the employment relationship is established, the legal protection available against their discriminatory potential is significantly restricted (in the case of Spanish law), even if under EU law, these preliminary stages of recruitment would be protected (cf. Firma Feryn case).

Meanwhile, automated emotion detection (psychogenic) systems are also applied in personal response systems based on bots which are being incorporated within the basic level of personnel management, and can be used to anticipate behaviours, providing a powerful tool for corporate control of workers.

The application of the profiling system applicable to applicants in the interchange between job offers and applications (as in the case of a public employment services) has thus become a known practice in

[6] The processing of personal data (big data) is defined as "any operation or set of operations carried out on personal data, such as collection, storage, conservation, alteration, retrieval, disclosure, making available, erasure, destruction or carrying out logical and/or arithmetic operations on those data" (Article 2(b) of Convention 108 of the Council of Europe for the Protection of Individuals with regard to Automatic Processing of Personal Data, 1981.

both the private and public spheres. In some European employment systems, then, we are beginning to see this being implemented as a candidate evaluation method in access to employment, as demonstrated by implementation within the Austrian employment system (Allhutter, Cech, Fischer, Grill & Mager, 2020)[7] or the recent introduction within the Spanish employment system (Send@), the public nature of which does not preclude a possible bias, even if mitigated in the design.

The algorithms handle data which characterise potential workers, retrieve information from relational and unstructured data, and formulate a set of recommendations (Protasiewicz, Pedrycz, Kozłowski, Dadas, Kopacz y Gałężewska, 2016). These operations can be performed as an entirely automated process or with human assistance, thanks to algorithms for optimisation, heuristics, artificial intelligence, machine learning, semantic data models and decision support systems (Protasiewicz, et al., 2016). Machine learning helps to link up the available data from various sources (data mining and tracking algorithms) to obtain the best choice in staff selection (recommendation algorithm), based on a combination of "semantic data" which interact with one another to identify personal/professional traits used to classify and evaluate people within a set of individuals (using three possible types of data or information: structured data, unstructured data, and information provided by the users or applicants in this case, Protasiewicz, et al., 2016).

In access to employment assisted by such tools, these techniques filter candidate applications, serving to automate much of the selection process, while at the same time retaining the human intervention demanded by Article 22 GDPR, as the tool performs an "assistance" role in the decision process (notwithstanding the computational capacity to direct the entire process up to issuance and notification of the decision adopted; cf. Elenius robot recruiter[8] or IPSoft's Amelia–).

4. Applications used for job evaluation and oversight

The simplification and efficiency delivered by automated decision systems leads on to the consideration of future applications for employment management, such as the choice of workers affected by a collective measure, the adjustment of hirings during peak production or service periods, or the management of occupational risk prevention in accordance with epidemiological predictions (as seen with the SARS-CoV-2 virus pandemic), among other functions. These are not all necessarily negative, since they could be used for the purpose of detecting historical discriminatory patterns: from occupational risk prevention to the analysis of the causality behind temporary contracts (and the adaptation of hiring in line with predicted needs), along with the detection of more beneficial conditions, as well as redundancies on economic, technical, organisational or production grounds.

Mathematical models for uses connected with labour relations have already been applied in the field of human resources, although their predictive potential is currently seeing a remarkable expansion. Such systems are thus already being used on digital service platforms to define worker remuneration, digital reputation and potential access to tasks (European Economic and Social

[7] The result of this implementation was criticised in comparative legal theory, since it ranks all jobseekers, resulting in very low employability for women based on a series of variables such as maternity, as a consequence of the biased design of the algorithm without any gender perspective (leading to a different evaluation of the impact of paternity and childcare in men and in women), as indicated by Frö hlich, Spiecker and Dö hmann (2018). The AMS algorithm of the Austrian public employment service has since January 2019 calculated the employment probabilities of unemployed people registered with the service, classifying them into different groups, with different training and employment expectations which shape the public job search support programmes, in accordance with various criteria which could encapsulate biases such as sex, age, home address and nationality. See https://www.ams.at/organisation/public-employment-service-austria/working--recruiting---studying.

[8] https://redelenius.com/. This is presented as a "quick and tireless" system, coding more than 36.9 billion data points per second, eliminating the risk of acquiring prejudices.

Committee, in its Opinion *Artificial Intelligence: anticipating its impact on work to ensure a fair transition*).

Here are just some of the uses of machine learning and other AI-based tools with potential in the world of employment: a) *causality of temporary contracts,* b) *organisation of work*, c) *worker evaluation and appraisal or assessment of effectiveness or performance,* d) *calculation of remuneration,* e) *occupational risk prevention*, f) *job control,* g) *termination of working relationships* (ILO, 2021), h) *detection of past discrimination at the company*, and i) *other purposes*, among which one could mention the detection of more beneficial conditions or the negotiation of a company's collective bargaining agreement.

One characteristic common to all of them, and which represents the focus of legal interest, is their opacity, since workers are deprived of access to the functional criteria applied to them, and have no means of knowing that the performance assessment systems (cf. Workplace Analytics) are not based on inclusive criteria, but define tasks and minutely break down the times dedicated to them, completely uninfluenced by any human and environmental factors that could have consequences for their position at the company and their conditions of employment, since they affect the points score they will receive. This ultimately leads us to such extreme situations as a failure to set aside times for personal needs, such as basic physiological functions and other factors (for which provision is made in Spanish law, in Article 4.4 of Act 10/2021, of 9 July 2021, on remote working, "within the operational capacity of the company in this sphere").

The fact is that the ultimate reason behind this paradigm shift lies in the capacity of AI-based technology in terms of monitoring and overseeing people, in this case workers. This new era has led to the normalisation of hyper-surveillance of work and workers ("the constant boss", Nguyen, 2021), even in the sphere of the courts. In Spain, the Supreme Court accepts the validity of geolocation of workers (Supreme Court Judgment 163/2021, of 8 February 2021) within the coordinates of time and work, with the sole condition that the cost and maintenance thereof is not passed on to the workers themselves. It likewise accepts surveillance by third parties by means of branding techniques, which measure customer satisfaction and include within corporate reputation mechanisms the evaluation of the work performed by its workers ("begging and bragging") while at the same time the workers themselves have become "prosumers", and also evaluators. They all end up participating in the digital evaluation of the service providers, thereby impacting on conditions of employment, or even their continuation at the company… incorporating their own biases and social prejudices, as one of the cogs within the management of the employment relationship… ultimately likewise conditioning the company's recruitment policy in response to customer tastes (a risk which is forbidden by EU law, according to the judgment in Firma-Feryn, C-54/07, of 10 July 2008).

In short, what is needed is a new, inclusive focus (Maitland, 2019: 150-159), which includes the supervision of the metrics employed in the different applications used in this sphere, as in the Bill passed by the State Legislature of California (Automated Decision Systems Accountability Act AB13, of 2021[9]).

[9] AB 13, 2021-2022 State Assemb., Reg. Ses. (7 December 2020), available at https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13. The City of New York had already in 2018 brought in the country's first algorithm accountability legislation, to supervise the use of algorithms on the part of government, to examine how error and bias are included within their design, and to recommend measures guaranteeing accuracy and fairness (passed into law on 11 December 2017, and available at https://www.nyclu.org/en/press-releases/city-council -pasa-primer-proyecto-de-ley-nació n-dirección n-transparencia-sesgo-uso-gubernamental).

III. DETECTION OF ALGORITHMIC DISCRIMINATION

1. Interrelations and conceptual equivalences between legal and computational language

Computational and artificial intelligence scientists have over time made mention of the existence of biases in artificial intelligence (AI), which they associated with various types, and linked to a holistic phenomenon incorporating various concepts, asserting that data-based tools will nonetheless always be much more precise than judgements issued by professionals (Grove et al., 2000).

The inclusion of jurists within this debate gives rise to an interesting synergy between the two spheres of knowledge, contributing the essential legal analysis of the discriminatory impact of algorithms. From this perspective, the translation from computational to legal language has even been addressed, with interrelations likewise seen as necessary both for the automation of the application of law and justice, and for the legal treatment of algorithmic biases. The fact is that, as Hildebrandt (2018) highlights, computational language could "erode the grammar and alphabet of modern positive law", which would require a new hermeneutic approach, demanding a proper understanding of the vocabulary and grammar of machine learning. In the analysis of the impact of biases, the specialist literature likewise indicates the divergences in language (Chouldechova, 2016, holds that the notion of indirect discrimination is not a statistical concept in the mathematical sense, but an ethical/legal concept, since a technical instrument free of predictive biases may cause indirect discrimination depending on the context to which it is applied), considering *fairness* to be equivalent to *absence of bias*.

Similarly, the concept of fairness used in the field of computation to analyse bias (which could more specifically correspond to equity) is far from aligned with the legal language, since within the context of anti-discrimination law other concepts in truth play the decisive role, as in the case of the balance between such impacts and the justifiability, proportionality and rationality of the decision, which are not strictly speaking equivalent to the principle of material justice in the application of the law. The European Parliament Resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law enforcement (2016/2225(INI)), refers in paragraph 22 (regarding the discriminatory impact of algorithm and dataset systems) to fairness as the principle which must prevail in the examination of predictions based on data analysis. As an instrument for the correction of legislation through adaptation to the particular circumstances of a specific case, the application thereof incorporating the criteria that inform the general principles of law, fairness means equality and individual treatment, and therefore requires that equal cases be treated equally, and unequal cases unequally, combining Aristotelian epieikeia and Roman aequitas as an "ethical/legal attribute of the law" (Ruiz-Gallardón, 2017).

In any event, there is an unquestionable need to align the two languages (Hildebrandt, 2021: 13), in order for computational science to incorporate the legal concepts of the context for which it is designed or in which it must act, or at least a shared legal language serving to adapt to different local or regional contexts. Nonetheless, there are issues derived from the inherent complexity of the law that make it difficult to establish a strict correspondence between algorithmic biases and discrimination in the legal sense, focused more on the outcome than the process, which is where the algorithmic bias lies (Vantin, 2021: 370; Foster, 2004, and Xenidis & Senden, 2020: 172) and in turn “they heighten the tensions already existing within the anti-discriminatory corpus of the EU” (Xenidis & Senden, 2020: 738).

The encroachment of third parties in the act of selection on which the decision-making process is based, together with the autonomy of machine learning, distorts both the form in which decisions are reached, and the operational method of the criteria employed for this purpose, as opposed to traditional "human" mechanisms, which are likewise not free of bias or arbitrariness. These may have human indications themselves superimposed (e.g. company instructions) for the design of the algorithm, or where applicable, those of the software developers, in which case they take on any programming or training biases that could have infected the model if it was not trained in a context

other than that in which it is required to operate. However, can biases derived from AI-based automated mechanisms be deemed equivalent to the legal concept of "discrimination"?

In the sphere of employment, according to the 1958 ILO Convention on discrimination (employment and occupation) (number 111), discrimination is "any distinction, exclusion or preference based on race, colour, sex, religion, political opinion, national extraction or social origin which has the effect of nullifying or impairing equality of opportunity or treatment in employment or occupation". If the concept corresponds to different treatment based on the personal characteristics of an individual, and is not tied to the intention of discrimination, but rather the outcome (Tomei, 2003) the (non-deliberate) bias derived from the algorithmic analysis may thus be seen as equivalent to the legal concept of discrimination, within the context of the aforementioned convention. If one compares this concept with that of the EU anti-discrimination directives applicable to the sphere of work (Directives 2000/78, 2000/46, and 2006/54), this confirms that similarly, whatever the scope of the Directive, it is identified as a situation in which "one person is treated less favourably than another is, has been or would be treated in a comparable situation, on any of the grounds referred to in Article 1" (of the respective Directive: in the case of 2000/78, "religion or belief, disability, age or sexual orientation as regards employment and occupation", within the context of Directive 2000/43, race or ethnic origin, and in the case Directive 2006/54, sex)[10]. As a result, without prejudice to delimitation of the scope of liability derived from discriminatory acts, the concept of discrimination is identified as different treatment based on a protected cause[11], *even if the conduct or act in question is not deliberate or the discriminatory outcome is not intended.*

This trait serves to cover all the effects derived from algorithmic biases, whether they are intentional biases, or accidental biases derived from data inference (including *proxy data*, invalidating the claimed neutrality offered by *anonymisation* of the data processed).

At the same time, it would serve to cover intersectional situations where the relevant trait for the algorithm is not a protected category, but another "secondary" trait if it appears or is associated with a trait that is covered, making provision for the lack of legal protection of intersectionality as a distinct legal category, to the extent that this could collaborate in or influence the biased outcome caused with regard to categories that are protected (e.g. the decisive trait for the algorithm could be overweight, but associated with a woman, a protected "category", this could have legal relevance as information for a potential gender-based discrimination claim). This entails affording a different legal value to that resulting from the algorithmic bias, because for the AI system, the specific value is based on the confluence of the influences (in this case, sex + appearance), while just one of the causes would probably have led to a different outcome, which might not be decisive in demonstrating sufficient evidence in order to bring a discrimination claim. By way of example, inferring from a postcode or home address, or name or other variable, whether a person belongs to a protected category involves combining different factors in the analysis, which taken separately could be "innocuous" or irrelevant, but which, acting in combination, together generate a result which, in the example considered, affects people of a particular national or ethnic origin, religion or economic status, such that this conclusion would determine the placement of the individual or individuals in a particular ranking, order or position, with detrimental effects in the sphere of employment (e.g. not being selected for a job). Meanwhile, accreditation of the evidence of discrimination becomes particularly difficult to the extent that it must involve finding the correlation between the data, which could be a "black box model" issue

[10] In Spain, the Draft Bill for genuine and effective equality of trans-people and for the guarantee of the rights of LGTBI people (https://www.igualdad.gob.es/servicios/participacion/audienciapublica/Documents/APL%20Igualdad%20Trans%20+LGTBI%20v4.pdf) also includes such protectable categories, not within the category of "sex", but as a separate category.

[11] Article 21 of the Charter of Fundamental Rights of the European Union: on grounds of sex, race, colour, ethnic origin, religion or belief.

(inherent to deep learning, where the reasons behind the result based on the big data fed into it are unknown [Mayson, 2019]), which means that the AI multiplies the bias, but does not help to identify it, since it is based on data correlations, when employing machine learning.

The data are thus the worst obstacle to providing evidence, even if it remains possible to employ traditional mechanisms for this purpose, in other words the existence of protected elements in the case, and the negative outcome for the claimant. Likewise, machine learning allows greater justifications to undermine the evidence (Xenidis & Senden, 2020: 747), since it operates in the field of indirect discrimination, which can thus be overcome through objective and reasonable justification. Meanwhile, the victim will in such cases be less aware of bias (Ebers, 2020: 79), the main barrier in many cases preventing the risk and its impact from becoming visible.

2. Insufficiency of anti-discrimination law

Various key concepts in this field of European Union law demonstrate the distortion brought in by an attempt to fit algorithmic discrimination within the framework of this discipline so as to fulfil the requirement for protection in such situations. This is the case with the concept of intersectionality itself, or forms of discrimination by association, by error, or what is known as multiple discrimination, when an attempt is made to apply this to decisions based on data inferences performed by artificial intelligence, since the correlation of characteristics and their interrelationships give rise to an outcome that is not equivalent to what humans might have arrived at, since the decision could have prioritised a confluent characteristic which might not have been noticed by the latter, and which they would similarly be unable to identify when validating the proposed decision offered by the algorithm. Above all, though, because identification of the cause of the discrimination could remain concealed for the purposes of activating such concepts, and the protection they offer. By way of example, the correlation of characteristics inferred by an algorithm could identify a connection between the individual evaluated and a protected condition as the basis of a negative decision (discrimination by association), but as this is not directly or perfectly visible, it could go unnoticed in the examination to detect discriminatory causes (e.g. the algorithm captures the connection between the individual evaluated and a neighbourhood association derived from the person's involvement in activities that it has organised, while at the same time detecting that most of these activities are connected with the Arab world, and hence inferring that the person belongs to this group, in order to exclude them from a job selection process), or worse still, that the confluent characteristics in such an analysis do not constitute protected status under EU law or national legislation, but did have a decisive influence in the conclusion reached by the AI system.

2.1. Multiple discrimination and intersectionality

Multiple discrimination forms part of the studies (European Commission, 2007 and 2009), programmatic texts and legislative initiatives of the European Union, aware of the prevalence of gender in this multi-causal phenomenon (Serra Cristóbal, 2020: 146), but this has not effectively been integrated within its *corpus iuris*, preventing it from being effectively applied. This is not the case with intersectional discrimination as a specific form of discrimination, where the final outcome prevails over the sum or multiplicity of causes (Makkonen, 2002).

Intersectional discrimination, a term coined by Crenshaw (1989), is based on the need specifically to consider situations of discrimination that function through the confluence of different factors, and lead to an outcome that is different from the mere sum of independent causes. In the digital sphere, the system of inferences carried out by machine learning may track different factors, not all of which are subject to specific protection under the laws of each country (as in the case of obesity, physical appearance, use of certain clothing, tattoos, piercings… a phenomenon known as "profiling", actions based on appearance, often including gender prejudices), the interrelationship of which leads to the end result, either through not being prioritised in a selection process, or a negative evaluation for other

purposes… Intersectional analysis, argues Serra Cristóbal (2020: 157)[12], "serves to analyse the interdependencies among various factors of oppression, while simultaneously promoting an indivisible and interdependent interpretation of human rights".

The lack of positive establishment of the concept thus far in European Union law[13] (with reference being made in paragraph 14 of Directive 2000/43 and paragraph 3 of Directive 2000/78, both in the preamble, without any specific linked content in the body of the regulation)[14] has left the effective protection of situations defined by the confluence of various grounds of discrimination bereft of legal mechanisms, as demonstrated by the ECJ Judgment of 24 November 2016, case C-443/15, *Parris*. This means that the interrelationship between computational and legal concepts lacks specific effects in practice, to the extent that no legal support exists, although the impact of algorithmic bias could be much greater, since the influences that machine learning might draw are capable of detecting features that would not be noted by humans and their social prejudices, avoiding their discriminatory impact, and as a consequence this form of discrimination reveals much more serious potential than straightforward or single-cause discrimination (Xenidis & Senden: 740), just as machine learning has greater capacity to give rise to discrimination by association, specifically because of its capacity to infer correlations among data (Wachter, 2020: 371). This discriminatory potential is further heightened if one takes into account its invisibility vis-à-vis anti-discrimination control tools, to which it may remain concealed or unnoticed. As a consequence, the refocusing of EU anti-discrimination law would unquestionably help to mitigate algorithmic biases, intensifying the urgency of dealing with an issue that has for too long been left pending resolution at the European level. One appropriate formula for this is that proposed by Wachter (2020: 371): the inclusion as protected legal categories of those individuals that are clearly related to protected elements. In procedural terms, the outright acceptance of this concept would underpin protection for victims of multiple and intersectional discrimination (likewise for discrimination caused by the use of AI systems), simplifying the contribution of one single element of evidence of intersectional bias, and hence facilitating the process of proving algorithmic bias (Vantin, 2021: 276).

2.2. Dysfunctions between the concepts of algorithmic bias and discrimination by association, by error, or multiple discrimination

Automated decisions may commit multiple discrimination as a consequence of the way they reproduce our complex social reality, with the intersectional convergence of different social prejudices, a confluence of which is liable to affect the same individuals (Xenidis, 2021: 739-740).

---

[12] See also Schiek, D. & Lawson, A. (dirs.), 2016, and Serra Cristóbal, R. (coord.), 2013.

[13] And also in the Spanish system, although the 2021 Proposal for an Comprehensive Equality Act of the Socialist Group does, in Article 6.3, cover intersectional discrimination "where different grounds set forth in this Act exist simultaneously or interact, generating a specific form of discrimination" (b). It likewise establishes that "in cases of multiple and intersectional discrimination, the motivation of the different treatment, under the terms of subsection 2 of Article 4, must arise in connection with each of the grounds of discrimination" (at https://www.congreso.es/public_oficiales/L14/CONG/BOCG/B/BOCG-14-B-146-1.PDF). Also, the Comprehensive Guarantee of Sexual Freedom Bill (approved by the Council Ministers on 6/7/2021, https://transparencia.gob.es/servicios-buscador/contenido/normaelaboracion.htm?id=NormaEV03L0-20200902&lang=es&fcAct=2021-06-30T12:23:27.739Z) includes in Article 2.5 the principle of the response to intersectional and multiple discrimination, defined as violence overlapping with other discriminatory factors,

[14] For Xenidis & Senden (2020: 738), the concept of multiple discrimination is present in Directive 2006/54. At the European level, the Judgment of the European Court of Human Rights, in the case B.S. v. Spain, of 24 July 2012, does apply the stated concept. Meanwhile, the 2008 proposal for a horizontal directive, not yet approved, 11531/08 SOC 411 JAI 368 MI 246-COM(2008) 426 final (consolidated text of 2017), specifically refers to the intersectionality and composition of grounds for discrimination among those protected by the Directives, with Article 2.2(a) prohibiting multiple discrimination.

Identifying and correcting this bias when it is derived from mathematical decision-making models increases the complexity of what is *per se* a cryptic task, since in computational terms it demands the combination of various corrective criteria so as to avoid segregating just one of the characteristics to be protected. In truth, though, the technique provides anti-discrimination law with access to options of much greater legal complexity, since it proves more accessible in designing a non-discrimination order (in other words to prioritise) regarding certain combined characteristics which coexist in one single individual (sex, ethnic origin, sexual orientation…) for an algorithm, than to respond with the necessary legal tools to a case of multiple discrimination. Simply because what is feasible in technical terms is not yet consolidated in law, since EU law[15] has not yet established positive provisions for a legal concept of multiple discrimination which would serve to identify and protect against this type of discriminatory intersection, rather than continuing to treat it as a sum of independent causes (Xenidis, 2021: 741). The conclusion that may be reached is thus that EU law is not in a position to respond to algorithmic discrimination, since it allows different manifestations of the impact of automated decisions to evade a narrowly defined legal framework which suffers from legal gaps, forcing the analysis of the scope or extent of discrimination to be confined to simplified parameters, with no room for multiple discrimination[16], nor to replace criteria associated with characteristics protected by EU anti-discrimination law (e.g. country of birth, in place of ethnic origin, as occurred in the Judgment of 6 April 2017, in case C-668/15, Jyske Finans A/S) [Gerards & Xanidis, 2020].

Furthermore, the concept of *discrimination by association* raises queries that remain unresolved. In practice, to the extent that association would not be identified with the correlation of proxy data, since this is simply an element that detects characteristics by proximity, but does not in itself function as an element of discrimination, it is not possible to sustain a "digital" concept of discrimination by association. One may, nonetheless, assert a clear connection between the two concepts, and considerable discriminatory potential derived from the greater precision that the association between the person subject to discrimination and the characteristic in question reveals in other persons within their immediate circle, which could be tracked by an AI system (e.g. relationship with a person with disability). Since correlations that might not be known to humans (but which, if known, could lead to discrimination) can more easily be captured by an automated system (through inferences in datasets), and hence may amplify the capacity of association for detrimental purposes (the clearest example, offered by the well-known Coleman judgment of 17 July 2008 [case C-303/06], is the algorithmic detection of the connection with a person with disability (direct relative) through different inferences, such as membership of an association of parents of people with functional diversity, or other types of data providing evidence that, although the company was unaware of this circumstance of their family life, the association in question does exist, as a feature feeding in to exclusion from the selection and recruitment process, or for the purposes of other corporate decisions).

In truth, the correlation of inferences among data involves interrelating traits that the algorithm or set of algorithms employed prioritise, or those that they discard, according to the optimal results provided

[15] The ECJ Judgment of 24 November 2016, in case C-443/15, *Parris*, rejects the consideration of two causes as multiple or intersectional discrimination if taken separately they are not discriminatory. In other words, it demands that they should be discriminatory in isolation, which could in truth be classified as the overlapping of causes, but not intersectionality with a greater negative impact associated with the combination of causes in a new outcome. According to the Proposal for a Comprehensive Equality Act promoted by the socialist Parliamentary group, intersectional discrimination is defined as the combination or interaction of various protected causes, giving rise to a specific form of discrimination (Article 6.3(b). Meanwhile, this text distinguishes the concept from multiple discrimination, assimilating it with the principal established by the aforementioned ECJ Judgment, as a situation in which "a person is discriminated simultaneously or consecutively because of two or more protected causes" (Article 6.3(a).

[16] FRA – European Union Agency for Fundamental Rights: *Inequalities and multiple discrimination in access to and quality of healthcare*, 2010, http://fra.europa.eu/en/publication/2013/inequalities-discrimination-healthcare.

by the analysis, which ultimately means that certain traits or characteristics are defined as more appropriate for the intended corporate purposes (e.g. employment), while others will be relegated. Whatever the case, in this automated operation it will be highly feasible to detect traits determining discrimination *by association* in the aforementioned sense.

Discrimination by error (based on incorrect consideration of the characteristics of the person or persons discriminated against) would seem *a priori* to be identified with errors in machine learning, in drawing inferences that attribute to certain traits consequences that might not be consistent with reality (errors derived from the logical rules, where there is an insufficient database, overrepresentation or underrepresentation of data categories which are in turn encoding the real world). The mental deductive process leading to false considerations may be human or computational, although the labelling of data to transform elements of reality into digital knowledge is human in nature, since it is performed by humans, who may be guilty of error and bias.

Now, the key question in this case is the authorship of the error in question, since in the human case this is clearly attributable to an incorrect understanding of the reality or of a specific trait of the person being evaluated or observed, while in the digital case this error has been committed by other intermediaries, and hence indirectly, or otherwise by the machine learning mechanism itself, which is ultimately also the responsibility of people. In the former case, the imputation of responsibility is subjective, while in the latter it may be objective if, at the employment law level, responsibility is attributed to the person employing the tools that give rise to the bias, if the employers are ultimately the subjects responsible for all effects of their legal relationship vis-à-vis those who are employed or "employable", derived from the regulation of corporate liability in the event of the involvement of third-party companies (subcontracting and transfer of enterprise), or with regard to the use of defective products within the context of occupational risk. In such cases, the contractual employment liability model overcomes the difficulties inherent to the opacity and problems in traceability of those responsible throughout the computational chain that led to the end result, whether through the design, the inputs, the labelling of the data, training in contexts other than those within which the system is created, or its application, while also identifying one main or sole party responsible (Vantin, 2021: 376).

3. Evaluation of automated decisions as discriminatory

3.1. Is it relevant to know how an algorithm acts in order legally to classify its impact as discriminatory?

The question raised is whether one really needs to know how an AI system arrived at a conclusion, or if we should concern ourselves solely with the outcome.

Viewed from this perspective, we could consider two response hypotheses: a) the *thesis of conduct*; b) the *thesis of outcome*. Our legal system prioritises the outcome, while also lending significance to conduct in itself, since this serves to determine the degree of seriousness of the offence and the resultant liability, although the mere existence of a discriminatory impact constitutes the necessary evidence in order, in the procedural context, to transfer the burden of evidence onto the perpetrator or perpetrators of the conduct that triggers the hypothetically discriminatory outcome. As a consequence, relevance must in fact be given to the process leading to the outcome and the factors involved in it, as well as an understanding of how these automated biases function, and how they should be classified from the legal perspective.

There is, then, a clear need for experts in computation and AI to be fully conversant with the legal terminology involved, in an attempt not only to find the correspondences between the two languages, but also to address the mitigation of biases with appropriate baggage, since data correlations are neither irrelevant nor neutral in legal terms.

3.2. Discriminatory impact of algorithms

In the light of the current use of AI systems with an impact on fundamental rights, it is essential to evaluate them from the perspective of equality and non-discrimination law, and to identify whether our legal system is in a position to offer a satisfactory response to claims of liability regarding automated decisions. This demands an analysis both of the nature of such forms of discrimination in the light of EU anti-discrimination directives (Directives 2000/78, 2000/43, and 2006/54), in addition to the possible classification as direct or indirect discrimination.

An analysis of the discriminatory impact of automated decisions demands that we hold them up against the rules of anti-discrimination law, which leads us to raise various questions, not all connected with the sphere of discrimination: a) can there be discriminatory intentionality in the use of an algorithm which offers biased conclusions?; b) does this need to exist, or can objective liability be attributed as a result of the intervention of automated mechanisms to reach decisions, or otherwise should they be deemed to be null and void because of a lack of human intervention?; c) does it constitute direct discrimination or indirect discrimination to base decisions on automation techniques with little human involvement, generating discriminatory biases?; d) are traditional safeguarding tools sufficient, or should we reinterpret these forms of discrimination, since they were designed to operate with other parameters?

According to Miné (2003: 5), "direct discrimination may be intentional and explicit with regard to the prohibited grounds", "but, as such discrimination is explicitly asserted, in particular in a regulation, with increasingly less frequency, the law places the emphasis on the outcome generated by the difference in treatment, according to an objective concept of discrimination", and "the intentional nature of discrimination no longer constitutes an essential element".

In the sphere of employment, the difference in treatment between two comparable situations remains relevant, over and above the purpose (economic, discriminatory or of some other kind). Such situations may or may not be shaped by the recourse to AI-based technical tools, but the comparable elements will be the situations themselves, not the nature of the actions addressing them (in this case, how the decision was reached). This being the case, these other aspects must be brought fully to bear in the evaluation of the hypothetically discriminatory behaviour.

The automated decision may correspond to the concept of either direct discrimination or indirect discrimination, since an algorithm may be designed with the strict purpose of ruling out certain characteristics, such as in the case of staff recruitment, or to conduct an evaluation based on apparently neutral criteria to the detriment of individuals with certain characteristics, or that deliberately disregard those which clearly shape the evaluation of their productivity (e.g. illness or disability), which could be classified as "non-inclusive design" of the algorithm. Whether this is a non-neutral practice or an apparently neutral use which is nonetheless liable to entail a particular disadvantage for individuals fulfilling one or more criteria (or would otherwise particularly represent a sex-based disadvantage for people, compared with those of the other sex), the fact is that the business decision ultimately constitutes discrimination. Unless the exception obviating the presumption of indirect discrimination would apply, in other words that the criterion or practice is objectively justified for a legitimate purpose, and the means employed are appropriate and necessary. In short, anti-discrimination law requires a reinterpretation in the light of these new forms of technological discrimination, serving to adjust the legal responses.

3.3. Classification as direct or indirect discrimination

Calibrating the discriminatory impact of decisions based on algorithms demands that we first examine whether the company in fact conveys and incorporates biased parameters within the algorithm when reaching such decisions. The response is not unequivocal, since the inherent design of the algorithm is built on the basis of instructions pursuing a purpose, and this is defined by those who order it to be

programmed, whether the businesses that will directly make use of it, or those that sell the corporate software to be purchased by employers in order to organise their "human resources". It is thus possible that the response may be negative (this would constitute explicit, direct discrimination, incorporated within the structure of the algorithm).

Now, if a machine learning algorithm functions as a black box, and simply evaluates all possible variables to reach the most accurate decision, the final (biased) decision could be contaminated, and constitute a case of indirect discrimination, without corresponding to the intentions of the employers that used it as the basis for a decision. Various questions arise:

a) Is, then, the ownership of the algorithm of relevance in giving rise to liability for discrimination? If one takes into account the irrelevance of intentionality in order to arrive at a classification as discriminatory, what would be the consequence of deliberately using biased architectures, vis-à-vis the mere acquisition of software which causes the same unintended outcome?
b) If the functional model of the algorithm employed, e.g. in staff recruitment, takes a historical pattern as its reference point, in other words the company's own biased background, would this then constitute a possible case of *unconscious direct discrimination*?

The key undoubtedly lies in the attitude of the company towards this risk: its passivity in the face of the possible perverse effect of the choice is what ultimately makes it complicit in the bias algorithm. And, of course, the prior actions which fed into and trained the algorithm, and which it reproduces as the ideal model to be followed, which were previously implemented by the company, because in this case, although it is not responsible for a direct order to discriminate (as would correspond to a deliberate design for this purpose), it is responsible for its past actions.

The question is of huge interest from the perspective of anti-discrimination law, in that it allows one to consider liability based on past behaviour when this constitutes the unconscious basis of new decisions suggested by a third party (the algorithm), but accepted by humans, as a consequence of the fact that an algorithm outside their decision-making sphere determined that they should be reproduced in order to assist in new decisions, if the company is unaware of how the machine learning functions, and its consequences (precisely because the algorithm is not motivated by the elements which serve as the basis to arrive at the conclusion). In this case, what truly matters will be its consent to validate the bias that the algorithm reproduces (confirming the mathematical model's proposal), and as a result, in employment terms, the employer remains responsible for its discriminatory action, and one may even consider whether its conduct constitutes direct rather than indirect discrimination (the conduct comprising direct discrimination may be based on an implicit but conscious mechanism, while indirect discrimination requires that the discriminatory impact should result in an effect above all on a group defined by a protected characteristic).

Likewise, the algorithm may replicate both direct and indirect past discrimination (*historical pattern*). Thus resulting in the effective displacement of the measures established in the company's equality plan to correct situations of gender-based discrimination as a consequence of the use of historical patterns for decision-making.

The legal analysis of the correlation between the company's reputational mechanisms in the evaluation of workers and business decisions is likewise of key importance in this approach, given the singular importance that the architecture of the algorithms employed should be free of parameters of a personal nature, such as health, disability, work-life balance or trade union rights, and also for the correction of the impact of a system for the evaluation of such characteristics. The fact is that algorithmic models for performance evaluation are systematically ignoring employment law criteria subject to particular legal protection against discrimination, as in the case of disability, or the need for work-life balance. The question may be exemplified in the Judgment of the Court of Bologna of 31 December 2020, number 29491, which positions this discriminatory trait specifically within the non-inclusive design, bereft of any corrective factor for situations which would serve to justify Glovo deliverers in cancelling their availability to fulfil a service, governed by the *Frank* algorithm, leading it

to exclude workers who are in such (legally justified) circumstances excluded from the call priority order, which would thus, by being repeated over time, even jeopardise their continuation in their job.

IV. PROTECTION THROUGH ANTI-DISCRIMINATION LAW

1. Regulatory frameworks addressing the opacity of automated decisions

Although business decisions generate direct responsibility on the part of employers, irrespective of how the decisions were built or advised, once the decision-making space in employment is occupied by algorithms, this alters the traditional response mechanisms, by modifying the defensive capacity of the workers, since their characteristics provide cover for the alleged objectivity of such decisions, presenting them as opaque and incontrovertible. In practice, the problem of opacity is a consequence in the field of corporate or public decision-making (Burrell, 2016). The mass use of data and the convergence of complexity and apparent objectivity ("cloudy illusion of objectivity" [Stoica, Riederer & Chaintreau, 2018]) raise a scenario of trust, diluting the perception of opacity, despite the tangible consequences of such decisions.

Much of the problem of opacity is centred both on the intangible and technical nature of the algorithms and their inherent legal configuration, in that they are subject to intellectual property rights. Both elements allow their direct users to maintain a hazy aura of opacity as to their decisions if they are automated. This barrier or opacity is challenged by the protective mechanisms derived from Regulation 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* [GDPR], and Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data, of 28 January 1981[17], both with regard to automated decision-making, until autonomous instruments are developed in this field (which will likewise not apply to the EU's future AI Act, begun in April 2021).

For this reason, the applicable legal framework (characterised by its central focus on personal data protection) demands an appropriate combination of the stated regulations, and anti-discrimination law. This means that while the former allow the motivational parameters of business decisions to be delimited, so as to appraise the validity of such decisions in accordance with parameters of transparency (including their *explicability*, or the right to an explanation as to their functioning [Goodman & Flaxman, 2017] and the possible exceptions, protected by intellectual property and corporate rights) and human intervention (per Article 22 GDPR), anti-discrimination law provides the rules for the distribution of the burden of evidence in order to examine in greater depth the hypothetically discriminatory nature of such decisions.

Aside from the rules delimiting obligations concerning personal data protection and the right to equality and non-discrimination (from which the duty to explain the automated decision is derived), the regulatory framework of intellectual property is also involved, and may bring in additional complexities in determining obligations and responsibilities, since it likewise entails problems of the competent legal forum, and the imputation or allocation of liability (Vantin, 2021: 371), however much such liabilities must, in the context of employment, be attributed at all times irrespective of the contractual relationship between companies and workers, the framework which defines the direct response by the former vis-à-vis the latter. However, this subdivision of the applicable regulations would, in order to achieve better alignment, require a reconsideration focused on protection against

[17] See Council of Europe (2017): Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, Committee of experts on internet intermediaries, MSI-NET(2016)06 rev6, https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a.

automated decisions, and in general the application of artificial intelligence to people, which would likewise cover personal data protection.

Notwithstanding the various strategies that may be ordered for the prevention and mitigation of bias, such as governance and algorithmic fairness, the risk impact assessment, auditing of algorithms and standardisation, supervision by trade unions and the public authority, technical mitigation, aside from the transparency of algorithms and other tools for intervention studied by specialists[18], such as those which could be covered by the Whistleblowing Directive (Directive (EU) 2019/1937 of the European Parliament and of the Council, of 20 October 2019, *on the protection of persons who report breaches of Union law*, and even the use of the data themselves in order to combat discriminatory biases (algorithmic fairness, Ho & Xiang, 2020)[19], the following analysis will focus on the strictly legal sphere from the perspective of discrimination, to which this seminar corresponds.

2. Access and explicability of algorithms

Algorithms are defined as "finite sequence(s) of formal rules (logical instructions and operations) allowing an outcome to be obtained from an information input", which means there are two crucial elements from a legal perspective: the sequence of instructions (the "source code" [20]) and the information or data that it uses (known as "libraries", or sets of data, which the legal analysis must in turn address with regard to ownership and rights of use, depending on their provenance, and access by any challenger), on which to focus the so-called right of *explicability* and transparency, equivalent to the motivation for the decision they help to assess.

2.1. Right of explicability and access to underlying reasoning

EU law links both core issues of protection: personal data and automated decisions, through the protection of personal data. The bridge leading to protection against discrimination is precisely the impact of data-based automated decisions, to the extent that these employ profiling (per Article 22 GDPR) and may contain biases, which are very difficult to understand, as a result of the inherent complexity of the machine learning on which they are based (Gunning, 2017). Hence the fact that one of the fundamental elements in structuring protection against algorithmic discrimination, until greater regulatory developments are achieved, would precisely be the analysis of data processing, while accepting the limitation of the scope of application and possibilities offered by Article 22 GDPR and, of course, Article 25 of the same text, which draws on tools which have revealed themselves to be entirely insufficient in the field of machine learning influences (pseudonymisation and other techniques to which said principle refers, to anonymise data or strip them of personal traits, since any type of trait is open to tracking and inference, which means that such techniques do not guarantee equality).

---

[18] See also Rivas Vallejo (dir.) (2022) for greater considerations in this regard.

[19] In the United States, the Obama administration (2016) already raised this need, resulting in the "Guidance for Regulation of Artificial Intelligence Applications" of 17 November 2020 (https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf), which proposes "bias mitigation". The UK Government's Race Disparity Unit gathers, analyses and publishes governmental data as to the experiences of people of different ethnic origins, in order to promote changes of policy where disparities exist (Centre for Data Ethics and Innovation [Blog], 2020).

[20] The source code may be defined as the "set of lines of text which are the instructions that must be followed by the computer to run the program; since it is in the source code that the computer's functionality is written", in alphanumeric characters, in a programming language chosen by programmers (such as: Basic, C, C++, C#, Java, Perl, Python, PHP). When applied to algorithms, "source code be understood as any human-readable text, written in a specific programming language. The purpose of the source code is to create clear rules and provisions for the computer, allowing it to translate them into its own language" (definition of the *Digital Guide, Ionos*).

At the European level, two legal instruments guarantee the right to data protection: the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Convention 108, and Regulation (EU) 2016/679 (GDPR).

With regard to the right of transparency and explicability, this may take the form, within the context of a discrimination claim, of objective and reasonable justification which may be demanded of the company responsible for the decision challenged, and which constitutes application of the distribution of the burden of proof in discrimination proceedings. Nonetheless, making the parameters and criteria of the decisions intelligible (*principle of algorithmic transparency*) proves difficult in such cases. The explicability of algorithms fulfils various functions: it helps to explain how they function, both for designers and developers, and for those impacted by their effects, likewise contributing to the reliability of the system that uses them (and the auditing of this system), and furthermore allows legal arguments to be built up in order to uncover their potential unlawfulness (Ebers, 2021: 48), or that of their impact, once applied in a particular context. It is therefore decisive to calibrate the degree of explicability and appraise the neutral exploration of the algorithms by a third party that remains outside the scope of intellectual property and business secrecy problems that could result from disclosure for the purposes of the search for biases (Ebers, 2021: 80). The auditing of algorithms becomes a key tool in this regard.

In the configuration of this right, the first significant question is the exclusion of protection of *non-personal* data, per Article 9.1(b) of Convention 108 (which guarantees the individual right to the communication of the data processed in an intelligible form, all information available as to their origin, the storage period, and any other information, so as to guarantee the transparency of processing), which allows for an exception: personal data that are not gathered from the data subjects, in which case the data controller is exempt from the obligation if the processing involves "disproportionate efforts". One may interpret that machine learning particularly complicates fulfilment of this duty, and such a situation could be deemed to constitute the "disproportionate efforts" referred to in Article 8.3.

Secondly, the legal regulation itself confines this to *online contracts* or *entirely automated processes*. In practice, data subjects are entitled to be informed of the motivation of the algorithm if it is used for profiling (the case governed by Article 22 GDPR, the purpose of which is to cover the public scope of data processing, for public order purposes, as indicated by the EU proposal for an AI Act), in other words to ascertain the *underlying rationale* in the data processing if the results thereof are applied to them (this is likewise the indication given by the opinion of the ESEC in referring to the fact that the principle of algorithmic transparency means making the parameters and criteria of the decisions taken intelligible), and that this explanation should be provided by humans. Recital 71 of the Regulation establishes this right of people accessing *services contracted online with no human involvement whatsoever*, and maintains that *this type of processing includes profiling comprising any form of personal data processing that evaluates personal aspects regarding a natural person* (trade union membership, employment performance, ethnic or racial origin, political opinions, religion or philosophical beliefs, data concerning health or sex life, criminal sentences and offences, or related security measures, Recital 75[21])*, in particular to analyse or predict aspects connected with employment performance… to the extent that this would produce legal effects concerning him or her, or similarly significantly affect him or her.* The regulation would, then, literally seem to delimit a restricted sphere of the right to explicability, which would thus not extend to automated decisions in part of the decision-making process, even if this is a particularly relevant part in the set of factors leading to the final decision, e.g. screening of CVs in a selection process, which is ultimately "humanised" by putting

[21] "Sensitive data" are, according to Article 6 of Convention 108, special categories of data protected by the aforementioned principle, which demand appropriate supplementary guarantees when they are processed, in particular racial or ethnic origin, political beliefs, trade union membership, religious or other beliefs, sex life or health… either independently or in combination with other data (*Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data*, 2015).

humans in charge of the decision, which, as occurred in the Uber case (Court of Amsterdam, judgment C/13/692003/HA RK 20-302), could have been confined to validating the recommendation of the AI system, which in turn may have used a historical pattern replicating previous biased decisions, or even if this is not the case, brings to an end a procedure in which the CV screening phase was entirely automated, even if the process involved humans in its final phase. In this example, if one considers, e.g., that a thousand CVs were submitted, of which fifteen were preselected for human examination, the bulk of the bias, where most of the discriminations based on various protected causes would probably lie, will have been included within an entirely automated phase, regarding which no formal explanation will be offered, when, as indicated in Recital 71, it does not involve "*online contractual services without any human intervention*", which could occur because they are not online contractual services, or because human intervention does exist within the decision-making chain. Hence the particular importance of delimiting the concept, to demand that the "human intervention" be "significant" or "substantial".

Thirdly, the contents of the explanation refer to the "underlying rationale", in other words the actual *reasoning mechanism of the algorithm* (the source code, or its modification through deep learning?), while the Spanish regulation only covers the right to be informed of the purpose of the processing, *not how it is technically implemented*. This raises a problem in making the use of machine learning devoid of content, since the decisive element in the decision will be the data. And even if the data are used for profiling, the scope of the right continues to be confined to this circumstance (and to the right to be informed of the right of objection, if the circumstances of Article 22 of the Regulation rise, which would thus likewise be within a restricted framework of application).

However, the right to basic information on the part of those competing in a selection process and affected by an algorithm which reaches the decision on the basis of analysis of data of third parties to which it furthermore does not extend the information already in its possession, since the former will probably already know the purpose of the data processing and the identity of those responsible (which, if not known, will likewise be of no great assistance in identifying the bias suffered), although the regulation ultimately adds that *basic information may in such cases include the data categories processed and the sources from which they are drawn*. In this case, the analysis of data categories would provide access to the source of the decision-making algorithm, but in a superficial manner, if the *source code* is not shared, and even if it is, it may likewise prove insufficient unless the data that fed the algorithm are shared (Huergo Lora, 2020: 54 ss., and Roig, 2020).

2.2. Significant human intervention

Human intervention connected with automated decisions has only one regulatory point of reference in our positive law: Article 22 GDPR (and the corresponding principle of domestic law). In the light of this principle, and its interpretation in the prior recitals, the right of explicability refers to *online contractual services without any human intervention*, thereby revealing the urgent need to extend this narrow sphere of legislative attention to the adoption of automated employment decisions. Meanwhile, however, even if one were to admit an interpretation extending to different spheres other than online contractual services, considering the use that companies make of different AI-supported tools to reach decisions (including staff selection and recruitment, not as an online or network process, although this may be used to deploy the selection process, or even simply to attract job applicants, e.g. CV screening), the concept of human intervention must be central to the corresponding regulation and legal analysis. In particular, the concept of human intervention, or human-in-command), likewise called for from the ethical and technological perspectives, requires a regulatory definition serving to offset enforceability in employment concepts. In other words, serving to exclude rubber-stamping of automated recommendations, or meaning that the human intervention referred to constitutes more than processing the confirmation order, and is instead substantial or significant in nature.

This idea may be observed from two perspectives:

a) That of *significant human intervention*. This must be capable of explaining that automated processing prior to the human decision involved ancillary aspects, or those that did not lead to a final, definitive decision for any individual. To employ a procedural comparison, this would refer to acts bringing the process in question to an end, e.g. access to a job (rather than to the process or procedure).

b) That of *definition of the concept of decision*. This case would involve including within the scope of guarantees not only those decisions that may be considered "final" or definitive, e.g. the hiring decision, but all those of relevance for the subjects involved, such as exclusion from the recruitment process, which would require the screening phase likewise to be supervised by humans, who would furthermore take responsibility for the process of notifying the parties concerned, thereby expressing the required *explicability*, and allowing the reason to be challenged.

In any event, one could argue that it would be desirable for both aspects of this approach to be included within automated decision-making models in the sphere of employment, since they are decisions that could affect fundamental rights, such as the right to equality and non-discrimination (which would require a hypothetical specific regulation of digital employment rights).

2.3. Access to motivation and intellectual property rights

Within the context of anti-discrimination protection, and in particular the objective and reasonable justification that may overcome the evidence of discrimination, as a consequence of the motivation of the decision challenged and the limitation of the scope of the right to human intervention (Article 22 GDPR), it may be essential to determine the extent to which access is possible to this intangible instrument on which the business decision was based (the sequencing comprising the algorithm), as well as the data fed into it (libraries or information in the form of big data). The incipient trend, however, is to focus claims against automated decisions on access to just one of these elements, the *source code*, in order to verify the defects in design that could establish the harm argued in the claim (and ascertain how the decision was reached).

The legal approach to the question must take into account two different situations: a) algorithms with a simpler architecture, not based on machine learning, in which the input data become secondary, as in the case raised against administrative applications for requests for public grants (in which the algorithm focuses on a binary [yes/no] decision [SCANTAMBURLO, 2021: 703]), and is only required to determine whether the applicants do or do not fulfil the criteria previously entered in designing the algorithm); or b) those based on machine learning or data input, as in the case of the predictive algorithms used in staff selection, where the data are specifically the key to the algorithm's learning in order to perform its prediction or selection (e.g. the Send@ algorithm of the Spanish State Public Employment Service). While in the former case an understanding of the source code or programming sequence would undoubtedly provide an insight into the origin of the decision (or its possible biased manipulation), in the latter case, such access will undoubtedly not provide the basis required to raise a cogent objection to the harm caused by the bias.

Now, the initial claims in the employment sphere have focused solely on the *source code* (cf. workers of Glovo[22] or Uber, requesting access to the "source code" of the algorithm), as a means of guaranteeing the right of transparency (although such access does not guarantee intelligibility, since it requires sufficient technical knowledge for its interpretation). However, if the patented software or

[22] In the first case, the Court of Amsterdam in its judgment C/13/692003/HA RK 20-302 of 11/3/2021 (and C/13/687315/HA RK 20-207 of the same date) ruled against them, in that the workers were provided with a sufficient explanation as to the functionality, and this was not challenged by them (without prejudice to the obligation to provide access to the personal data of those affected). The second case was resolved in the judgment of the Ordinary Court of Bologna on 27/11/2020, which did find that the algorithms used by the company were not inclusive.

algorithms are the intellectual property of their creators (or those who acquired them, if the rights were assigned), it is possible that the employer using the software does not enjoy ownership nor any right of disposal whatsoever (e.g. companies acquiring a user licence) that would allow the right of transparency to be put into practice.

The employer could challenge such a claim by citing intellectual property rights in the algorithm, applicable to "the various parts of a work… provided that they contain elements which are the expression of the intellectual creation of the author" (ECJ Judgment, Infopaq International, C-5/08, Rec. p. I-6569, paragraph 39, and ECJ Judgment, Grand Chamber, of 2 May 2012, in the case SAS Institute Inc v. World Programming Ltd., paragraph 65). This protection does not extend to the ideas and principles on which the elements of a computer program are based, including those which underlie its interfaces, according to Directive 2009/24/EC, of the European Parliament and of the Council, of 23 April 2009, on the legal protection of computer programs, Recital 11 of which explicitly refers to algorithms: "in accordance with this principle of copyright, to the extent that logic, algorithms and programming languages comprise ideas and principles, those ideas and principles are not protected under this Directive", and must be protected "by copyright" under national legislation. Such protection, according to Recital 63 of the GDPR, allows the source code to be concealed in the context of a claim (likewise, mere observation, study or verification of the functioning of a program, without prior authorisation of the proprietor, does not constitute a legal breach, under Article 5.3 of Directive 2009/24/EC, but in the light of EU case-law, this mere observation cannot be identified as access to the source code).

Along similar lines, if the company obtained a licensed copy of the algorithm in question, then under Directive 2009/24/EC, it would likewise be authorised to "observe, study or test the functioning of the program, provided that those acts do not infringe the copyright in the program", because they are not protected by the copyright covered by the Directive (ECJ Judgment, Grand Chamber, of 2 May 2012, in the case SAS Institute Inc v. World Programming Ltd, paragraph 50). Similarly, paragraph 61 of the judgment identifies such a situation as the mere use of the program, without access to the source code, distinguishing between this and studying, observing and testing..., before concluding that "keywords, syntax, commands and combinations of commands, options, defaults and iterations consist of words, figures or mathematical concepts which, considered in isolation, are not, as such, an intellectual creation of the author of the computer program" (paragraph 66), although it is "only through the choice, sequence and combination of those words, figures or mathematical concepts that the author may express his creativity in an original manner and achieve a result", the computer program user manual, which is an intellectual creation (ECJ Judgment of 16 July 2009, Infopaq International, C-5/08, Rec. p. I-6569, paragraph 39). In conclusion, although the decision refers to another core of analysis (the copy of a computer program or a part thereof), the court finds that this combination which we call an algorithm does constitute an intellectual creation, which is documented and written in code language, since "the object of the protection conferred by that directive is the expression in any form of a computer program which permits reproduction in different computer languages, such as the source code and the object code" (ECJ Judgment of 22 December 2010, *Bezpecnostní softwarová asociase case*, paragraph 35).

3. Evidence and proof of algorithmic discrimination

3.1. Accreditation of evidence of discrimination in the case of algorithmic biases

In order to allow an appraisal of the discriminatory nature of a decision, one necessary condition is accreditation of the evidence of discrimination, in accordance with the rules of evidence (Article 8 Directive 2000/43/CE, Article 10 Directive 2000/78/CE and Article 19 Directive 2006/54/EC), although it is difficult to answer the question as to whether decision-making by means of intervening digital mechanisms does or does not make it more difficult to prove evidence.

As a starting hypothesis, one should not *a priori* rule out that the interpretation of the appearance of discrimination or sufficient evidence (Carrizosa, 2012: 59-65) -*prima facie*- would not be altered as a consequence of the use of an algorithm, since they are specifically presented as tools for objectivity and precision in advising on decisions. On a scenario such as a worker selection process, the bias of the algorithm, if it exists, may be accredited by means of traditional evidentiary mechanisms (e.g. comparison between the individuals chosen, and those excluded). Now, once the evidence has been accepted, and taking into account the great precision with which the selection algorithm can operate, unless the bias is to be found in its inherent design (source code), how does one perform the comparative analysis within the universe processed by it, with regard to the individual excluded from the selection? In other words, if in a selection process which involves hundreds of candidates, groups of different individuals with various protected (or even non-protected) characteristics are excluded, the comparison between the claimant and those not ruled out will not serve to confirm that the reason for exclusion was solely the characteristics possessed by that individual, since a whole host of inferred characteristics may likewise have been rejected.

With regard to the binary male/female distinction, and the systematic exclusion of women, the analysis would seem simple, but if the case involves rejection of characteristics corresponding to different groups (disability, ethnicity, religion, origin, among others) but not others, or not their combination with others, the examination involved becomes complex, since the multitude of variables combined by the automated model also allow one to consider that other characteristics that may have been identified could have been prioritised, but could have gone unnoticed in accreditation of the required evidence. It is even more challenging to classify decisions as discriminatory if the criteria employed by the algorithm do not precisely correlate to characteristics attributable to a protected social category, although they could likewise be related to such a characteristic. For example, taking as our reference the case cited by Xenidis (2021: 4), if the basis of the algorithm's decision lies in variables such as distance from the workplace, this does not *per se* determine the existence of a discriminatory criteria under positive law, but if that factor is combined with a particular origin, which is connected with a protected characteristic, the inference performed by the algorithm results in a discriminatory decision (e.g. the individuals that live in the rejected area are mainly from an immigrant population). If, in turn, the characteristic with which the association is made is not one of those protected by anti-discrimination law, a situation of intersectional discrimination could arise, as a result of the coexistence of various factors that only a more comprehensive analysis could reveal, and that only a more extensive framework of protection than that provided by the directives could classify as discriminatory. Now, it is also important to bear in mind the possible severance of the nexus of connection required in order to establish the necessary relationship between the decision and the claimed discrimination, precisely because of the distance between the reference data and the consequence analysed. In short, what matters most is that the inherent functional dynamic of the automated decision mechanisms makes it hugely difficult to detect discriminatory biases, and thus demands a refinement of the granularity of the analysis, and reconsideration of the legal strategy as regards protection against discrimination.

3.2. In the case of multiple and/or intersectional discrimination

Multiple and intersectional discrimination find a direct equivalent between the methods of data inference and the discriminatory outcome, which means that the legal acceptance of the autonomy of both concepts would serve to provide fair coverage for pockets of discrimination which, through the effect of the automated mechanisms studied, may not only be subject to an exponential increase which is to a great extent invisible, but furthermore also lie outside the scope of adequate legal protection.

Continuing to the level of evidence, it is of interest to analyse whether the equivalence between intersectionality and the bias resulting from the inference between data fields is open to evidentiary accreditation. To the extent that machine learning does not allow us to know the correlation between

data, and which of them determined the outcome, in other words it is not possible to know which of the traits analysed were decisive in the results offered up by the automated mechanism, and if it is not technically feasible to provide this explanation to the hypothetical victim of the discriminatory decision, we will most likely be faced with a case of multiple, intersectional discrimination, without it being possible to confirm this, unless the employer also provides the comparative data (not the input data). In short, this is the traditional system for the presentation of evidence, facilitated by the use of another automated system capable of finding the correlation between a set of individuals and the individual claiming the discrimination, which could also be assisted by automated mechanisms capable of detecting the simple or intersectional statistical impact of discriminatory traits.

Bibliography

ALLHUTTER, Doris, CECH, Florian, FISCHER, Fabian, GRILL, Gabriel & MAGER, Astrid (2020): "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective". *Front big data*. DOI: 10.3389/fdata.2020.00005.

ARAGÜEZ VALENZUELA, Lucía (2021): "Los algoritmos digitales en el trabajo. Brechas y sesgos". *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*. Volume 9, issue 4. ADAPT University Press.

BAROCAS, Solon & SELBST, Andrew D. (2016): "Big data's disparate impact". *California Law Review*, issue 104, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

BERSIN, Josh (2019): "The Skills Of The Future Are Now Clear: And Despite What You Think, They're Not Technical". Author's blog, https://joshbersin.com/2019/09/the-skills-of-the-future-are-now-clear-and-despite-what-you-think-theyre-not-technical/ (8/9/2019).

BUCHER, Taina (2018): *If... Then: Algorithmic Power and Politics*. Oxford: Oxford University Press, DOI: 10.1093 / oso / 9780190493028.001.0001.

BURRELL, Jenna (2016): "How the machine 'thinks': understanding opacity in machine learning algorithms". Vol. 3, issue 1, https://doi.org/10.1177/2053951715622512.

CARRIZOSA, Esther (2012): "La concreción de los indicios de discriminación en la jurisprudencia comunitaria: STJUE 19 abril 2012", *Aranzadi Social*, vol. 5, issue 7, pp. 59-65.

CHOULDECHOVA, Alexandra (2016): "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments", pp. 1-17, en https://arxiv.org/abs/1610.07524.

CRENSHAW, Kimberle (1989): "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics". *The University of Chicago Legal Forum: 1989*. HeinOnline - 1989 U. Chi. Legal F. 139, https://philpapers.org/archive/CREDTI.pdf?ncid=txtlnkusaolp00000603.

EBERS, Martin (2020): "Ethical and legal challenges", in EBERS, Martin, & NAVAS, Susana, dirs. (2020): *Algorithms and law*. Cambridge University Press.

FLORES, Anthony W., BECHTEL, Kristin, & LOWENKAMP, Christopher T. (2016): "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias': There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks". *Federal Probation*. Volume 80, issue 2.

GERARDS, Janneke & XENIDIS, Raphaële (2020): "Algorithmic discrimination in Europe: Challenges and Opportunities for EU equality law", *European Futures*, 3/12/2020, https://www.europeanfutures.ed.ac.uk/algorithmic-discrimination-in-europe-challenges-and-opportunities-for-eu-equality-law/.

GILLESPIE, Tarleton (2016): "Algorithm". In *Digital Keywords: A Vocabulary of Information Society and Culture*, ed B. Peters (Princeton, NJ: Princeton University Press), doi: 10.1515/9781400880553-004, y https://www.researchgate.net/publication/309964434_2_Algorithm.

GROVE, William M., ZALD, David H., LEBOW, Boyd S., SNITZ, Beth E. & NELSON, Chad (2000): "Clinical versus mechanical prediction: a meta-analysis". *Psychological Assessment*, vol. 12, issue 1.

GUNNING, David (2017): "Explainable Artificial Intelligence (XAI)", https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

HAJIAN, Sara, BONCHI, Francesco, & CASTILLO, Carlos (2016): "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining", DOI: 10.1145/2939672.2945386, https://www.researchgate.net/publication/305997939_Algorithmic_Bias_From_Discrimination_Discovery_to_Fairness-aware_Data_Mining.

HILDEBRANDT, Mireille (2018): "Algorithmic regulation and the rule of law". *Philosophical Transactions of the Royal Society A*, vol. 376, issue 2128. DOI:http://dx.doi.org/10.1098/rsta.2017.0355.

HILDEBRANDT, Mireille (2021): "The issue of bias. The framing powers of ML". *Computer Science*, DOI:10.2139/ssrn.3497597. In M. Pelillo, T. Scantamburlo (eds.): *Machine We Trust. Perspectives on Dependable AI*, MIT Press 2021, http://dx.doi.org/10.2139/ssrn.3497597. Preprint.

HO, Daniel E., & XIANG, Alice (2020): "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness". *The University of Chicago Law Review Online* (30/10/2020), https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang/.

HUERGO LORA, Alejandro (2020): "Una aproximación a los algoritmos desde el Derecho administrativo", in HUERGO LORA, Alejandro (dir.) & DÍAZ GONZÁLEZ, Gustavo Manuel: *La regulación de los algoritmos*. Aranzadi, Cizur Menor.

LAAKSONEN, Salla-Maaria, HAAPOJA, Jesse, KINNUNEN, Teemu, NELIMARKKA, Matti, & PÖYHTÄRI, Reeta (2020): "The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring". *Front. Big Data*, 5/2/2020, https://doi.org/10.3389/fdata.2020.00003.

MACKENZIE, Adrian (2017): *Machine Learners: Archaeology of Data Practice*. Cambridge, MA: The MIT Press.

MAKKONEN, Timo (2002): Multiple, Compound and Intersectional Discrimination: bringing the experiences of the most marginalized to the fore. Institute For Human Rights, Abo Akademi University.

MAYER-SCHÖNBERGER, Viktor & CUKIER, Kenneth (2013): *Big Data.* Turner Publicaciones, Madrid. In http://catedradatos.com.ar/media/3.-Big-data.-La-revolucion-de-los-datos-masivos-Noema-Spanish-Edition-Viktor-Mayer-Schonberger-Kenneth-Cukier.pdf.

MAYSON, Sandra G. (2019): "Bias In, Bias Out". The Yale Law Journal, vol. 128, issue 8, https://www.yalelawjournal.org/article/bias-in-bias-out#:~:text=abstract.,to%20have%20disparate%20racial%20impacts.

MINÉ, Michel (2003): "Los conceptos de discriminación directa e indirecta", Conference: "Lucha contra la discriminación: Las nuevas directivas de 2000 sobre la igualdad de trato", 31/3-1/4/2003, Trier, http://www.era-comm.eu/oldoku/Adiskri/02_Key_concepts/2003_Mine_ES.pdf.

NGUYEN, Aiha (2021): *The Constant Boss, Work Under Digital Surveillance*. Data & Society, in https://datasociety.net/wp-content/uploads/2021/05/The_Constant_Boss.pdf.

O'NEIL, Cathy (2017): *Armas de destrucción matemática*. Capitán Swing, Madrid.

PASQUALE, Frank (2019): "A Rule of Persons, Not Machines: The Limits of Legal Automation", The George Washington Law Review, vol. 87, issue 1, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3135549.

PASQUALE, Frank (2020): New laws of robotics: defending human expertise in the age of AI. The Belknap Press.

PROTASIEWICZ, Jarosław, PEDRYCZ, Witold, KOZŁOWSKI, Marek, DADAS, Sławomir, STANISŁAWEK, Tomasz, KOPACZ, Agata, & GAŁĘŻEWSKA, Małgorzata (2016): "A recommender system of reviewers and experts in reviewing problems". *Knowledge-Based Systems*, vol. 106, pp. 1643178, DOI: 10.1016 j.knosys.2016.05.041.

PUJOL VILA, Oriol (2021): "The concept of 'artificial intelligence'. Opactity and societal impact". 2020. In Pablo García Mexía & Francisco Pérez Bes (eds.): *Artificial Intelligence and the law*. Wolters Kluwer.

RIVAS VALLEJO, Pilar (2020): La aplicación de la inteligencia artificial al trabajo y su impacto discirminatorio. Thomson Reuters Aranzadi, Cizur Menor.

RIVAS VALLEJO, Pilar, dir. (2022): Discriminación algorítmica en el ámbito laboral: perspectiva de género e intervención. Thomson Reuters Aranzadi, Cizur Menor.

ROIG, Antoni (2020): *Las garantías frente a las decisiones automatizadas: del Reglamento General de Protección de Datos a la gobernanza algorítmica*. Bosch, Barcelona.

ROSENBLAT, Álex (2018): Uberland Cómo los algoritmos están reescribiendo las reglas de trabajo. University of California Press.

RUIZ-GALLARDÓN, Isabel (2017): "La equidad: una justicia más justa". *Foro, Nueva época*, vol. 20, issue 2, pp. 173-191, http://dx.doi.org/10.5209/FORO.59013.

SCANTAMBURLO, Teresa (2021): "Non-empirical problems in fair machine learning". *Ethics and Information Technology*, issue 23, pp.703–712, https://doi.org/10.1007/s10676-021-09608-9.

SCHIEK, Dagmar y LAWSON, Anna (dirs.) (2016): European Union Non-Discrimination Law and Intersectionality: Investigating the triangle of racial, gender and disability discrimination, London-New York: Routledge.

SERRA CRISTÓBAL, Rosario (coord.) (2013): La discriminación múltiple en los ordenamientos jurídicos español y europeo, Valencia: Tirant lo Blanch.

SERRA CRISTÓBAL, Rosario (2020): "El reconocimiento de la discriminación múltiple por los tribunales". *Teoría y derecho*, issue 27, pp. 140-161. DOI: https://doi.org/10.36151/td.2020.008.

SLAVIN, Kevin (2011): "Cómo los algoritmos configuran nuestro mundo", TED talks, 11/7/2011, https://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world?language=es.

SMITH-STROTHER, Lisa (2016): "The role of social advocacy in diversity & inclusion recruiting", Glassdoor Summit 2016, https://youtu.be/IdsqQMV4V_0).

SPIEGELHALTER, David, & HARFORD, Tim (2014): "Big data: are we making a big mistake?" *The Financial Times*, 28/3/2014, en https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

STOICA, Ana-Andreea, RIEDERER, Christopher, y CHAINTREAU, Augustin (2018): "Algorithmic glass ceiling in social networks: the effects of social recommendations on network diversity". *Proceedings of the Web Conference* 2018, Lyon. ACM, Nueva York, pp. 923–932, https://doi.org/10.1145/3178876.3186140.

TOMEI, Manuela (2003): "Análisis de los conceptos de discriminación y de igualdad en el trabajo". *Revista Internacional del Trabajo*, vol. 122, issue 4.

VANTIN, Serena (2021): "Inteligencia artificial y derecho antidiscriminatorio", en LLANO ALONSO, F. & GARRIDO MARTÍN, J. (eds.): *Inteligencia artificial y derecho. El jurista ante los retos de la era digital*. Thomson Reuters Aranzadi, Cizur Menor.

XENIDIS, Raphaële (2021): "Tuning EU equality law to algorithmic discrimination: three pathways to resilience", *Maastricht Journal of European and Comparative Law* 2020, vol. 27, issue 6, 4/1/2021, en https://doi.org/10.1177/1023263X20982173.

XENIDIS, Raphaële y SENDEN, Linda (2020): "EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination". U. Bernitz et al. (eds.): *General principles of EU law and the EU digital order*. Kluwer Law Int., pp. 151-182.

WACHTER, Sandra (2020): "Affinity profiling and discrimination by association in online behavioural advertising". *Berkeley Technology Law Journal*, issue 35, https://btlj.org/data/articles2020/35_2/01-Wachter_WEB_03-25-21.pdf.

ZUIDERVEEN BORGESIUS, Frederik (2018): *Discrimination, artificial intelligence, and algorithmic decision-making*. Council of Europe, Directorate General of Democracy.

NOTE: all digital links were verified in the month of January 2022.