Artificial Intelligence-based Discrimination: Theoretical and Normative Responses. Perspectives from Europe

di Costanza Nardocci

Abstract: La discriminazione che deriva dall'Intelligenza artificiale: la teoria e le risposte del diritto positivo. Prospettive a partire dall'esperienza europea - The paper examines the relationships between AI and discrimination. The first part challenges the adequacy of anti-discrimination laws to tackle AI-based discrimination. The second analyses the regulatory responses proposed by the European Union and the Council of Europe. The investigation includes a study of the case law, which highlights the challenges prompted by AI when coupled with the principle of non-discrimination and the lack of effective legislative and judicial remedies to counter it. Lastly, the paper argues that AI is contributing to the emergence of a new form of discrimination, the global acknowledgment of which is still far away in coming.

Keywords: Artificial Intelligence; Non-Discrimination; Victims; Proxy; Proxy Discrimination.

1. Introduction

Artificial intelligence is becoming increasingly pervasive in everyday life and in the realm of human rights. Within this scenario, the principles of equality and non-discrimination are shown as being interconnected with artificial intelligence (hereinafter AI) with the latter starting to be considered as being highly likely responsible for discriminatory conducts in a variety of sectors.

Despite the undeniable link between the two phenomena at stake, in that AI might cause discrimination, the law is resistant to regulating technological innovations brought about by AI and to depicting discrimination as one of the major challenges AI might generate.

The paper sets out the specifics of AI-based discrimination, which deserves an autonomous definition from human-driven discrimination (Part I) to then zoom in on some tentative normative responses to AI and human rights examining the current *statuses* attributed to AI in the European continent, particularly in the European Union and the Council of Europe (Part II). Additionally, Part II includes the, sad to say, few cases of AI-based discrimination successfully brought before national Courts to show the role the judiciary might play in reacting to AI suspected of generating discrimination.

The overall aim is to highlight the criticisms and the negative impact that the poor theoretical analysis surrounding AI-based discrimination, which is most likely to be compared to instead of being countered by traditional forms of discrimination, has in the law and the role of the judiciary to attain a firm grasp and sanction properly AI-based discrimination.

Part I. The Theory. "Based" or "Artificial" a New Type of Discrimination

1. Making Differences vs. Discriminating: Humans vs. the "Machines"

> *"Part of the challenge of understanding algorithmic*
> *oppression is to understand that mathematical formulations*
> *to drive automated decisions made by human*
> *beings"*[1]

There is something wrong or lacking when discrimination approaches AI and *vice-versa*. The two phenomena happen to be very close to one another, as the latter quite often causes or at least poses risks of discrimination, but how it does so is considerably different from discrimination arising from human conduct[2].

From a theoretical viewpoint, keeping in mind who or what is the original, unique, or even partial agent in causing discrimination is of paramount importance to fully understand: first, how AI functions in a potentially discriminatory manner; second, to disclose how the law and the judiciary might intervene to tackle AI-based discrimination.

It is widely known that the European and US anti-discrimination laws[3] are based on the existence of a causal link between the (human) conduct and the discriminatory effect.

It is also commonly acknowledged that humans are the "solo" actors of the discriminatory conducts, which could lead to easily prove their liability, especially in cases of direct discriminations or disparate treatments.

On the contrary, the intertwined relationship, featuring the connection between humans and the machine in AI and the discrimination deriving from this, represents the preliminary and undeniable distinction

[1] S.U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, 2018.

[2] On this, see M. Favaretto, E. De Clercq, B. Simone Elger, *Big data and discrimination: perils, promises and solutions. A systematic review*, in *Journal of Big Data*, 2019, 1 ff. With regard to the differences existing as a result of the mediation of the machine in the discriminatory functioning of AI systems, see A. Simoncini, *L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà*, in *BioLaw Journal – Rivista di BioDiritto*, no. 1/2019, 63 ff.

[3] In European Union law, reference is made to M. BELL, *Anti-Discrimination and the European Union*, Oxford, 2002; C. Mccrudden, *Anti-Discrimination Law*, Aldershot, Dartmouth, 2004; D. Schiek, V. Chege, *European Union Non-Discrimination Law. Comparative perspectives on multidimensional equality*, London, 2009; S. Fredman, *Discrimiantion Law*, Oxford, 2011.

that separates AI-based discrimination from classical human-driven discrimination[4]. In other words, in these cases it takes two to tango.

Where liability lies and who is responsible for AI-based discrimination is, therefore, complicated when: the origin of the conduct is unknown; the discrimination constitutes a mixture of interconnected actions performed by the programmer (usually, the human) and the machine; and, also when the machine itself differentiates without any reasonable and/or objective justification.

As well as the cases when differentiating results in a discrimination, it should be highlighted that making a distinction does not in itself and always mean to discriminate against someone. AI technologies are actually built to make decisions and to distinguish between facts, words, and faces, without for this simple reason implying that such options should be sanctioned as discriminatory by the law[5].

The separation between humans and AI reflects the one existing between humans and the machines[6]. Unlike a machine, a human being is almost always fully aware and in control of what they do. Put differently, humans possess what is called metacognition, or the ability to distinguish bad from good, to identify the bad and to learn from the mistakes made. AI does not possess any sort of metacognition.

The inherent differences between humans and AI are of course mirrored in how both actors discriminate. While it is true that humans and AI systems are likewise capable of discriminating, nevertheless, at the same time there are unavoidable differences in how humans and AI act when confronted with making choices. This will entail an attempt to adapt or at least reconsider the adequacy of existing anti-discrimination laws to effectively counter discriminations based on the functioning of AI systems.

2. How and Why AI Discriminates: the Conduct

AI might be the cause of discriminations in a variety of ways.

It could be driven by human action as when it is the human that makes use of the machine with the sole purpose of discriminating against someone

[4] On the specifics of AI-derived discrimination, see F.Z. Burgesius, *Discrimination, artificial intelligence, and algorithmic decision-making*, Council of Europe Publications, 2018; S. Barocas, A.D. Selbst, *Big data disparate impact*, in *California Law Review*, 2016, 671 ff.; J. Kleinberg, J. Ludwig, S. Mullainathan, C.R. Sunstein, *Discrimination in the Age of Algorithms*, in *Journal of Legal Analysis*, 2018, 113 ff., and, of the same A., also, J. Kleinberg, J. Ludwig, S. Mullainathan, A. Rambachan, *Algorithmic fairness*, in *AEA papers and proceedings*, 2018, 22 ff.; see, also, C. Nardocci, *Intelligenza artificiale e discriminazioni*, in *Rivista "Gruppo di Pisa*, 2021, link: https://www.gruppodipisa.it/images/rivista/pdf/Costanza_Nardocci_-_Intelligenza_artificiale_e_discriminazioni.pdf.

[5] K. Lippert-Rasmussen, *Born free and equal? A philosophical inquiry into the nature of discrimination*, Oxford, 2014.

[6] On this, see S.M. Fleming, *What separates humans from AI? It's doubt*, *Financial Times*, 26 April 2021. In line with this argument, also, A. Rouvroy, *The end(s) of critique: Data-behaviourism vs. due-process*, in M. Hildebrandt, K. de Vries (eds.), *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, Milton Park and New York, 2013, 143 ff

and/or a group. This phenomenon is called "masking"[7] to emphasize that the liability lies only with the human whereas the machine merely performs as a medium to discriminate with.

Beyond the masking, which does not pose significant challenges to anti-discrimination laws, there are other ways in which AI discriminates. It happens sometimes that AI unreasonably discriminates with the involuntary complicity of the human who is implicitly biased against a category, and, in other circumstances, it appears that AI itself is the one and only agent to cause discrimination.

While discriminatory effects are easily traceable and sometimes even rapidly unveiled when the human is responsible, identifying the process leading to AI-based discrimination is instead challenging, as the agent of the discrimination is often hidden and the cause is likewise not recognizable on an *ex-post* basis.

More importantly, the discriminatory process, meaning the link between the conduct and the effects, diverges from that of classical human discrimination[8]. The reason lies in the fact that the machine plays a more or less prominent role in discriminating and the conduct could result as being the product of a complex relationship between the human, the programmer and the machine or, the AI system. This is why by reconciling and understanding what causes discrimination when AI comes into play might result in an ever-ending exercise, trying to find who (if it is the human) or what (if it is the machine), or more likely both is involved in the discriminatory act at stake.

The awareness of the interplay between humans and machines does not help in solving the dilemma of how AI discriminates. One may wonder which of the two actors plays a prominent role, and to what extent the programmer is capable of influencing and, later especially, controlling and supervising the functional abilities of the machine.

In short, AI-based discrimination first challenges anti-discrimination law, because the difference in treatment cannot be attributed to the unique responsibility of a human being, but rather to the intricate relation between AI and human beings. Moreover, even time matters in this case. Depending on "when" the human intervenes in the process could impact how AI-based discrimination performs, which forms it takes (on this, see below), and which victims will most likely be affected by it.

If the conduct at the origin of AI-based discrimination is unknown, the same argument could be used when trying to discover the causal link that lies between action and effects. However, the traits of the conduct in AI-based discrimination are hardly the only "new" features, nor do they complete the extensive list of questions surrounding the new form of discrimination.

[7] On this notion, please refer to the studies of L.J. Strahilevitz, *Privacy versus Antidiscrimination*, in *Chiacago Law Review*, 2008, 363 ff.; K. Lippert-Rasmussen, *Statistical (And Non-Statistical) Discrimination*, in *The Routledge Handbook of the Ethics of Discrimination*, cit.

[8] More extensively, please, refer to C. Nardocci, *Intelligenza artificiale e discriminazioni*, in *Rivista del "Gruppo di Pisa"*, 2021.

The following aspect looks at how AI discriminates. As well as the dichotomy between those who minimize AI-based discrimination to a solo-human prejudice problem and those who consider it a technical adversarial effect of technologies whose functioning should be rectified, the existing legal literature tends to primarily recall the role of data and their selection. Instead of being neutral, data are conceived as inherently human with the consequence that the more data are humanly originated, the more it is understandable that AI will be capable of discriminating. In other words, the same prejudices and stereotypes occurring in the human realm, that reflect past or systemic discrimination, are incorporated into the data which will likewise replicate those same human biases[9].

Whereas it is tangible that data are somehow a human product, the lack of neutrality of the data does not explain how AI ends up discriminating. At least five steps or moments should be looked at in the attempt to understand AI-based discrimination[10].

The first, mentioned already, is the selection of the data fed into the machine. During this phase, bias might act as one of the key factors leading to discrimination. The heterogeneity of human biases widely affects the quality of the data and their representation of the outside reality in cases of over or under-representation of certain groups or categories, which are usually those already suffering from structural discriminations[11].

The second step or phase is the training of the data, which could be autonomous run or, conversely, guided by the programmer. Depending on the type of AI technologies, this phase may, therefore, show a more or less determinant influence on the programmer: the more AI functions as a machine learning system, the less the programmer will be able to control and supervise the outcome of the technology at stake. Put differently, the machine will take precedence over the human in the possible discriminatory outcome of the machine.

The training of the data is also relevant as it could witness two interesting phenomena: the poisoning of the data and the inaccurate or not updated information provided to the machine in the light of scientific and technological innovation. Data poisoning occurs when the human – and here the liability lies almost entirely with the programmer – deliberately feeds the machine with data that are poisoned, unhealthy, untruthful and

[9] On this, see, extensively, K. Crawford, *Think Again: Big Data*, 2013, link: http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data; of the same A., also, *The Hidden Biases in Big Data*, in *Harvard Business Review*, 2013. For a proposal of categorization of *bias*, see S. Quintarelli, F. Corea, F. Fossa, A. Loreggia, S. Sapienza, *AI: profili etici. Una prospettiva etica sull'intelligenza artificiale: principi, diritti e raccomandazioni*, in *BioLaw Journal*, 2019, 218 ff. On this, also, H. Suresh, J.V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning*, MIT, 2020; in the Italian literature, reference is made to M. D'Amico, *Una parità ambigua. Costituzione e diritti delle donne*, Raffaello Cortina Editore, Milan, 2020 and, more recently, of the same Author, *Parole che separano. Linguaggio, Costituzione, diritti*, Milan, 2023.

[10] Reference is made to S. Barocas, A.D. Selbst, *Big data disparate impact*, in *California Law Review*, 2016, 671 ff.; F.Z. Burgesius, *Discrimination, artificial intelligence, and algorithmic decision-making*, Strasbourg, 2018, 10 ff. In the Italian literature, see P. Zuddas, *Intelligenza artificiale e discriminazioni*, in in *Giurcost.*, 2020, 1 ff.

[11] See J. Lerman, *Big Data and Its Exclusions*, in *Stanford Law Review Online*, 2013.

misleading. Instead, the latter case concerns AI technologies whose dataset has not been updated in a way consistent with the advancement and development of innovation, therefore impairing AI's ability to rightly respond to the tasks assigned.

Alongside these first two phases, there are additional phases where discrimination infiltrates AI. Literature[12] speaks of the crucial momentum of the identification and selection of the "target variables" and "class labels", which are used to group into categories; the "feature selection", meaning the choice of the features used by AI; more fundamentally, the choice of the "proxy" as the element AI will refer to in order to make distinctions which may eventually turn out to be of a discriminatory nature. All the above mechanisms, taken alone or in conjunction with one another, represent an attempt to cause the biased functioning of AI systems.

Moreover and along with the technical phases mentioned above, AI might discriminate due to the automaticity of its functioning. Not every distinction amounts to discrimination, but the failure of the machine to be capable of recognizing and subsequently treating equally two analogous situations, or differently unequal cases, and, lastly, of uncovering a reasonable justification for differentiating similar cases greatly increase the risks of biases and discriminatory outcomes of AI technologies. Furthermore, the opacity of these systems, according to the notorious "black-box theory"[13], is an additional reason for an explaination for AI-based discrimination.

Such complexity should suggest two preliminary conclusions. First, the conduct that lies behind AI-based discrimination is widely different from purely human discrimination. Second, as it will be examined in Part Two, said heterogeneity suggests that legislators and Courts should adjust existing anti-discrimination laws in the light of the specifics of this "new" discrimination. Notwithstanding the opacity and peculiarities of this "new" discrimination, human causes should not be underestimated that continue to rely on existing power relationships among human beings, stereotypes, and prejudices against the most vulnerable groups.

2.1. The Deficiency of Direct Discrimination towards (Statistical) Indirect Discrimination

The first and most significant challenge posed by AI-based discrimination to the classical categories of anti-discrimination law deals with direct discrimination (disparate treatment)[14].

[12] On this, extensively, S. Barocas, A.D. Selbst, *Big data disparate impact*, in *California Law Review*, 2016, 671, ff.

[13] See extensively F. Pasquale, *The Black Box Society. The secret algorithms that control money and information*, Harvard University Press, Cambridge, 2015. The identification of strategies to counter the theory at issue to foster an "explainable AI (xAI)", see A. Deeks, *The judicial demand for explainable artificial intelligence*, in *Columbia Law Review*, 2019, 1829 ff., and D. Lehr, P. Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, *Davis Law Review*, 2017, 653 ff.

[14] Broadly, argue the deficiencies of traditional anti-discrimination law to capture the specifics of AI-derived discrimination in the national literature, see J. Gerards,

As is widely known, direct discrimination requires a discriminatory intent, hinges on one or more explicit suspected grounds, and creates a more or less clear-cut unreasonable distinction between two similar or comparable situations. All these elements should also be supported by proof of the recurrence of a causal link between the conduct and the discriminatory effects on the victim's side.

Whereas in human-driven discrimination, the disparate treatment is exclusively caused by the action of a human being, the conduct in AI-based discrimination results from the connection between the human and the machine. Put differently, the causal link between the act and the discriminatory effects is complicated by the role played by AI in the former, which also affects how the difference in treatment will operate depending on the type of AI technology at stake.

As a consequence, AI might: a) influence the already discriminatory conduct of the human behind the machine; b) contribute with a more or less pervasive impact to discriminate together with the human being; c) entirely neglect the agency of the programmer, causing the discriminatory effect itself alone. The latter case could be that of machine learning or deep learning systems if endorsed by the so-called "black-box" theory which would deny any liability on the side of the programmer.

Nevertheless, all cases demonstrate that proving the existence of the causal link and the intentionality of the discriminatory conduct is negatively affected by the often unknown functioning of AI. The difficulty to define the relationship between the human and the machine in discriminating thus implies the inapplicability of the disparate treatment theory in finding the violation of the principle of equality. Moreover, it is very unlikely that AI makes distinctions explicitly grounded on prohibited factors of discrimination as required by definition in a direct discrimination case, relying instead on proxies. In resorting to the proxy despite a factor of discrimination, AI-based discrimination shows again the unfeasibility of AI-based discrimination as a classical form of direct discrimination.

Likewise, the structure of indirect discrimination hardly reconciles with AI-based discrimination. Besides the apparent neutrality, which could assimilate AI-based discrimination into an example of disparate impact, there is still an element that contradicts such a statement.

First, and once again, the distinction would not be grounded on the traditional factors of discrimination. The correlation between the proxy used

Algorithmic discrimination in Europe: Challenges and Opportunities for EU equality law, consultabile al link: https://www.europeanfutures.ed.ac.uk/algorithmic-discrimination-in-europe-challenges-and-opportunities-for-eu-equality-law/. L. Giacomelli, *Big brother is «gendering» you. Il diritto antidiscriminatorio alla prova dell'intelligenza artificiale: quale tutela per il corpo digitale?*, in *BioLaw Journal*, 2019, 269 ff; P. Zuddas, *Intelligenza artificiale e discriminazioni*, in *Giurcost.*, 2020, 1 ff.; M. Lees, *The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union*, in *Security Dialogue*, 2014, 494 ff. See, also, CoE's CAHAI, l'*Ad hoc Committee on Artificial Intelligence*, in its working paper *Feasibility study on a legal framework on AI design, development and application based on CoE standards*, December 17th, 2020.

by AI and the suspect ground, which needs to be proved even in a classical disparate impact case, is often unclear, confused, or, in the worst-case scenario, the proxy itself is unknown. Moreover, correlations are unpredictable which complicates the feasibility of *ex-ante* strategies to tackle AI-based discrimination.

Intentionality, moving on, could be missing. The endorsement of doctrines that believe proved indirect discrimination even when the intent to discriminate on the part of the agent is missing does not alter the heterogeneity of AI-based discrimination when compared to indirect discrimination.

More broadly, the opacity of AI greatly affects the *ex-post* evaluation of the features of the conduct and, at the outset its classification as a) direct or b) indirect discrimination. Additionally, the obscure functioning of the machine overshadows the identification of the intent, impairing the proof of the recurrence of direct discrimination. Once more, the factor of discrimination and its identification is lacking.

In short, and particularly referring to the latter, the more the criteria of the distinction are hidden and unknown, the more difficult it will be to discern who is or might be the comparator. Without a comparator[15], be it concrete or hypothetical, there is no discrimination, neither direct nor indirect.

2.2. Proxy and Unconscious AI-Based Discrimination

Much has been said about the failure to include AI-based discrimination in the categories of anti-discrimination law.

Conversely, less has been said about what the main features of AI-based discrimination are: what this "new" type of difference in treatment is; what its causes are; what the elements are that characterize its external manifestations. Whereas it might be demonstrated and it could be true that AI itself operates in a discriminatory manner, causing unreasonable disparate treatments or impacts, at the top of its functioning there are always the humans who originally program the operational systems.

This is extremely important because if there is something in common between AI-based discrimination and human-driven discrimination, it is the entire human cause of discrimination. Prejudice and stereotypes (understood as unreasonably and automatic categorizations), meaning biases, are at the core of both phenomena despite the subsequent differences that, as said, prove the inadequacy of anti-discrimination laws to respond to the challenges brought about by AI-based discrimination.

Bias and the lack of neutrality of the algorithm depict one form of AI-based discrimination. Reference is made to unconscious disparate treatment and the unconscious disparate impact that are types of AI-based

[15] On the challenges brought up by the difficulties in identifying the comparator, see E. Lundberg, *Automated decision-making vs indirect discrimination. Solution or aggravation?*, in https://www.diva-portal.org/smash/get/diva2:1331907/FULLTEXT01.pdf.

discrimination, which rely on implicit biases, meaning prejudice and stereotypes the programmers are unaware of[16].

This type of discrimination is exactly the opposite of the "masking" mentioned above, in that all the phases where AI could be at risk of resulting in discriminatory treatment are the product of original unconscious biases inherently rooted in the person who creates the machine. In other words, intentionality is often lacking when it comes down to AI-based discrimination. The absence of an explicit intent and its combination with the inexplicable functioning of the machine represents relevant factors in proving the liability of the agent (the human being).

A second type or way of describing the specifics of AI-based discrimination revolves around the proxy[17]. AI makes distinctions motivated by elements that have – maybe merely apparently – nothing to do with suspected grounds of discrimination. A system could make choices based on a zip code, which may seem not to be in any relationship with human qualities favoring discriminatory effects. On the contrary, the proxy could be associated with an extremely wide range of a factor of discrimination, acting exactly as one of them.

The problem with proxy discrimination deriving from AI rests exactly on the legality of the chosen criteria of distinction which, at least at first sight, renders AI free from any form of scrutiny and sanction for its non-conformity with the principles of equality and non-discrimination[18].

Nevertheless, the proxy does not merely act as a synonym of a factor of discrimination, in that being associated with one or more of the suspected grounds, it favors the discrimination of already disadvantaged groups. It also "creates" new suspected grounds of discrimination, new elements that, because of their links with the traditional factor of discrimination, generate discriminatory effects sometimes directly – "direct proxy discrimination"[19] – sometimes indirectly – "indirect proxy discrimination".

Without dwelling any further on these two types of discrimination (unconscious AI discrimination and proxy discrimination), it could nevertheless be pointed out that: a) these two phenomena represent the more

[16] On the notion and role played by implicit biases in AI's discriminatory functioning, see C. Jolls, C.R. Sunstein, *The Law of Implicit Bias*, in *California Law Review*, 2006, 969 ff.; N. Schmid, B. Stephens, *An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination*, in *ArXiv*, 2019, 130 ff. See, also, H. Suresh, J.V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning*, MIT, 2020.

[17] On the role of the proxy within the data-sets, see B.A. Williams, C.F. Brooks, Y. Shmargad, *How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications*, in *Journal of Information Policy*, 2018, 78 ff.; A. Datta et al., *Proxy Discrimination in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs*, 2017, in https://arxiv. org/pdf/1707.08120.pdf.

[18] To explore the notion and implications of proxy discrimination, see A.E.R. Prince, D. Schwarcz, *Proxy discrimination in the age of artificial intelligence and big data*, in *Iowa Law Journal*, 1277 ff.; A. Datta et al., *Proxy Discrimination in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs*, 2017, in https://arxiv. org/pdf/1707.08120.pdf.; and, also, J. Grimmelmann, D. Westreich, *Incomprehensible Discrimination*, in *California Law Review*, 2017, 164 ff.

[19] L. Alexander, K. Cole, *Discrimination by Proxy*, in *Constitutional Commentary*, 1997, 453 ff.

frequent ways in which AI manifests its discriminatory attitudes towards human beings; and b) that legislators and Courts should become more and more aware of the heterogeneity of AI-based discrimination by properly taking into account its specifics distancing it from the classical purely human-driven discrimination.

3. The "Who" and the "What". Individuals, Groups, Sub-Groups: from Old to New and Unaware Victims and from Old to New Identification Traits

The argued specificity of AI-based discrimination not only relies on the conduct of the agent, be it the human together with the machine. This also emerges while looking at the other side of the phenomenon, meaning taking into account the victim of AI-based discrimination and the identification trait that lies behind the decision made by the machine.

As widely known, discrimination is by definition a social phenomenon that possesses a deep-rooted collective nature, in that it originates from power relationships and conflicts among social groups[20]. The legal understanding of discrimination and the liberal approach entrenched in national constitutions and international treaties has later led to the interpretation of discrimination as a violation of the principle of equality affecting the individual first and foremost.

Although this dichotomy between the collective and individual dimension of discrimination is only apparent, AI-based discrimination has contributed to unraveling and highlighting the first dimension of discrimination, meaning its negative impact on groups and sub-groups[21].

[20] On this aspect, see D.L. Horowitz, *Ethnic groups in conflict*, University of California Press, Oakland, 1985; M.N. Marger, *Race and Ethnic Relations. American and Global Perspectives*, Wadsworth Cengage Learning, Boston, 2009; A.D. Smith, *The Ethnic Revival*, Cambridge, 1981 and, also, of the same Author, *The Ethnic Origins of Nations*, Oxford, 1988.

[21] Examples are widely known. Women, first. On this, see the notorious study of J Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Also, and with regard to the non-gender neutral implications and functioning of AI, see the Reports issued by the World Economic Forum in 2018 (https://reports.weforum.org/global-gender-gap-report-2018/assessing-gender-gaps-in-artificial-intelligence/?doing_wp_cron=1621003660.5886778831481933593750.) and by UNESCO in 2020 (https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000374174&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_ab07646d-c784-4a4e-96a1-3be7855b6f76%3F_%3D374174eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000374174/PDF/374174eng.pdf#AI%20Gender_pages.indd%3A.11061%3A142.). On this, also, M. D'Amico, *Una parità ambigua. Costituzione e diritti delle donne*, cit.; R. Adams, N.N. Loideain, *Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law*, in *Annual Cambridge International Law Conference New Technologies: New Challenges for Democracy and International Law*, 2019, 1 ff.; S. Leavy, *Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning*, in *ACM/IEEE 1st International*

This is an obvious consequence of how AI functions. The associations realized by the machine to make choices necessary rest on factors that identify groups or sub-groups rather than single individuals. This leads us to think that the victim of AI-based discrimination are certainly individuals, but individuals grouped as part of the category chosen or excluded by the machine.

The problem here is not so much the drive to reconsider the collective dimension of discrimination theoretically[22], which is also common to other forms of human-driven discrimination (e.g. those regarding cultural rights), but instead to identify to consider that the identification of the victim of AI-based discrimination will require to be based on some sort of criteria of belonging to a group or a sub-group of the individual personally affected.

The establishment of such belonging is nevertheless complicated by the fact that AI seldom makes a distinction based on traditional factors of discrimination but instead on the proxy and the proxy could be anything.

The extremely varied and completely unpredictable links between the proxy and the traditional factors of discrimination in some cases cause the proxy to act by affecting already disadvantaged social groups (e.g. racial minorities), and in others by endangering the individual who is entirely unaware of their association with the targeted group.

In other words, AI discriminates already marginalized categories but, at the same time, it creates "new" out-groups and, therefore, "new", but more fundamentally, unaware minorities[23].

All the results from the key and leading role of the proxy could be regarded as the elements that more than others characterize AI-based discrimination. It fosters new individual affiliations; additional and diverse types of identification traits, that are not included among the traditional factors of discrimination; new individual and group victims. The outcome is sometimes the impossibility for the victim to acknowledge that they have been discriminated against by AI and, even more so, to be part of a targeted group.

From a constitutional standpoint, the unawareness of the individual of being considered as belonging to a social group represents a serious violation of the constitutional principle of self-determination as well as of

Workshop on Gender Equality in Software Engineering, 2018, 14 ff. Another very well-known example is represented by racial and ethnic minorities. For an investigation on this, see, among others, R.M. O'Donnell, *Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause*, in *New York University Law Review*, 2019, 545 ff.; with regard to example of racial biases in AI systems, see Z. Obermeyer et al., *Dissecting racial bias in an algorithm used to manage the health of populations*, in *Science*, 2019, 447 ff.; C. Intachomphoo, O.D. Gundersen, *Artificial Intelligence and Race: a Systematic Review*, in *Legal Information Management*, 2020, 74 ff.

[22] With this regard, reference could be made to C. Nardocci, *Razza e etnia. La discriminazione tra individuo e gruppo nella dimensione costituzionale e sovranazionale*, Napoli, 2016.

[23] See B. Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, in *Philosophy and Technology*, 2017, 475 ff.; S. Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising*, in *Berkeley Technology Law Journal*, 2020.

the principle of self-identification laid down by the Council of Europe's Framework Convention on National Minorities under Article 3, § 1[24].

In the light of the above, it could be argued that by hinging on the proxy, discrimination derived from AI fosters the creation of "new" vulnerable groups, and, at the same time, of similarly "new" identification traits.

"New" victims have something in common: that of being regrouped by means of an element of division among human beings, which is powerful as a traditional factor of discrimination and which acts separating one group from another despite a. being associated with a factor of discrimination, without being itself a factor of discrimination; b. often being hidden and unknown; and, c. the absolute unawareness of the individual of its ascribed membership to the targeted group, which impedes the rights to access to justice of the victim, at least until they realize that they have suffered from discrimination.

Yet, it will be for the law to identify mechanisms to counter this new form of discrimination, to assist the victims and to safeguard their fundamental rights to equality and non-discrimination.

Even more so, it should be recalled that the new categorizations boosted by AI-based discrimination, by affecting new groups sharing unprecedented and new human features, aggravate and eventually impair the ability of the law to intercept the negative effects of AI systems especially when affecting both and sometimes even simultaneously "old", meaning traditionally disadvantaged groups, and "new", "unaware", victims of discrimination arising from AI.

Part II. Laws and Courts: Beyond Anti-Discrimination Laws to Tackle Inequalities in AI

1. A Space for Regulation: AI-Discrimination in the Law

AI has been largely absent from legal debates and, to an even greater extent, from parliamentary discussions.

In recent years, following a number of groundbreaking studies showing the severe risks and human rights implications of AI technologies, it has become the center of intense debate as to the advisability of regulating its functioning.

The dichotomy between self-regulation strategies and normative approaches has for a long time characterized the two sides of the Atlantic Ocean. The European continent has almost always projected its interest in AI endorsing the view that technological innovation should be embedded into the law, as a possible reflection of the civil law nature of the majority of its countries. While on the other side the United States especially has shown

[24] On this, for a better understanding of the linkages existing between AI and its discriminatory impact on the minorities in light of the CoE's Framework Convention see H.J. Heintze, *Article 3*, in M. Weller (ed.), *Oxford Commentaries on International Law. The Rights of Minorities. A Commentary on the European Framework Convention for the Protection of National Minorities*, Oxford, 2005, 124 ff.

first of all disregard, and secondly denial of intervention, regarding self-regulation as the best mechanism innovation that could control and rightly self-govern itself.

Such a construction of the relationships between Anglo-Saxon countries and Europe has become less and less a true representation of the reality in at least the last two years, bearing witness to the fact that it is still testifying a growing attempt to regulate AI even where the principle of self-regulation seemed to be insurmountable. It suffices to think about the "EU-US TTC Joint Roadmap", aimed at fostering cooperation between Europe and the United States on trustworthy AI, but also the "AI Bill of Rights" proposed by the President of the United States, Joe Biden, or the more recent "Strategic Enforcement Plan ('SEP')" of the Equal Employment Opportunity Commission ("EEOC") specifically tailored to tackle AI-based discrimination in employment, along with some States' initiatives such as the 2023 "Stop Discrimination by Algorithms Act" adopted in the District of Columbia (DC).

As well as Europe and North America, now China is the first and leading country in regulating AI systems followed by Japan, thus overtaking the European Union which was, conversely, at the forefront in April 2021 when it presented its proposal for a regulation of AI technologies (the so-called Artificial Intelligence Act).

The following paragraphs will examine the proposed EU regulation, the ongoing debate and criticisms also in the light of the prominent role of the Council of Europe (CoE), which recently published the draft of a first Framework Convention on AI relying on a robust human rights-based approach.

2. A Tentative Normative Approach: the EU Artificial Intelligence ACT (AIA) and the Lack of anti-Discrimination Policies

For some time now, the European Union has been regarded as the leading example in the regulatory approach. Be it for the civil law systems that are predominant in the EU, be it for the emerging continental fear of the negative implications of a certain type of AI, the EU presented its first proposal for regulation in April 2021[25]. At that time, the EU was the first international organization worldwide to declare its willingness to legislate on AI and to define sets of rules applicable within its Member States.

Despite the apparent rapid drafting process that preceded the publication of the original text, almost two years later the EU is currently struggling with the outline of a consensual regulation, with alternative outcomes roughly every six months of the presidency.

Not surprisingly, from being the first, the EU has been more recently replaced by China that in instead 2022 adopted the very first global example of norms governing AI.

[25] See first version of the text at the following link: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

The EU's 2021 text, *Proposal for a Regulation laying down harmonised rules on artificial intelligence*[26], aimed to move forward with the GDPR (*General Data Protection Regulation*) and its exclusive privacy-based approach[27], subjecting AIA to a heterogeneous set of norms resting on "risk criteria" to categorize AI technologies into groups, ranging from prohibited systems to acceptable ones[28].

Furthermore, the proposal includes some guiding principles for the regulation of AI to strengthen the connection between AI and human rights: transparency, human oversight, fairness, and explainability[29]. In order to comply with EU law, the algorithm should, thus, first be transparent as to the ways in which it is programmed and developed. Humans are always required to be kept in the loop, avoiding the risks advanced by the black box theory, and favoring human control, management and supervision of AI technologies. Moreover, explainability should be interpreted as a corollary of the principle of transparency, in that there cannot be the former without the latter.

Lastly, but in line with the scope of the analysis, the algorithm must be fair. AI should not be biased or, even implicitly, endorse a lack of impartiality that features its human origin. Fairness is, perhaps, the most significant bond between AI and discrimination in the document.

Apart from these guidelines and the obligations to develop AI in compliance with the principles provided under the EU Charter of Fundamental Rights, equality and non-discrimination are practically missing and poorly recalled in the text. The word "discrimination" is mentioned roughly twice and the only reference to the likelihood that AI

[26] The preliminary text might be read at the following link: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

[27] On the gaps of the GDPR in terms of the challenges prompted by AI in the perspective of human rights and, especially, with regard to the principle of equality and non-discrimination, see, among others, M. Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond*, in *International Journal of Law and Information*, 2019, 91 ff.

[28] See Titles II and III respectively dedicated to: *Prohibited Artificial Intelligence Practices* and *High-Risk AI Systems.*

[29] On the notion of explainability of an AI systems, please refer to S. Wachter, B. Mittelstadt, L. Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, 2017, who underline that: "[t]wo kinds of explanations may be in question, depending on whether one refers to: system functionality, that is, the logic, significance, envisaged consequences and general functionality of an automated decision-making system, e.g. the system's requirements specification, decision trees, pre-defined models, criteria, and classification structures; or to specific decisions, that is, the rationale, reasons, and individual circumstances of a specific automated decision, e.g. the weighting of features, machine-defined case- specific decision rules, information about reference or profile groups. Furthermore, one can also distinguish between explanations in terms of their timing in relation to the decision-making process: an ex ante explanation occurs prior to an automated decision-making taking place. Note that an ex ante explanation can logically address only system functionality, as the rationale of a specific decision cannot be known before the decision is made; an ex post explanation occurs after an automated decision has taken place. Note that an ex post explanation can address both system functionality and the rationale of a specific decision".

might result in unreasonable differentiations among human beings and social groups is limited to the concept of profiling.

Interestingly, in 2022, the AIA was subject to some amendment proposals by three EU Commissions during the legislative process before the EU Parliament, which followed the publication of a new compromised version of the text under the Slovenian Presidency. The two amendment proposals zoomed in on two prominent aspects: the drive to reconcile the AIA with EU anti-discrimination law and the need to contrast the limited sensibility of the text with the human rights implications and negative impacts of AI technologies.

The ITRE Commission and the JURI Commission, for instance, clarify that the governance and the management of AI datasets should not be confined to a mere investigation of the general lack of bias, but will have to proceed to an accurate analysis to certify the absence of risks of any sort for the fundamental rights provided under EU Law. Similarly, worthy of mention is Article 29*b*, *Fundamental rights impact assessment for high-risk AI systems*, of the ITRE Commission which suggests, that high-risk technologies should be subject to preventive scrutiny of their conformity with human rights.

Despite a trend to endorse a human rights-based approach to AI, it is both noteworthy and controversial is that neither the new version of the AIA nor the Commissions make references to the EU Directive on anti-discrimination law. It thus remains unclear whether the EU institutions are truly inclined to include discrimination and equality within the scope of the AI regulatory framework or if it is instead more widely concerned with AI's human rights without delving into and rooting out the real-life consequences of the unfair discriminatory functioning of AI technologies.

A more sensitive and structured consideration of human rights coupled with AI is vice versa the main goal of the statement submitted in November 2021 by almost 114 NGOs[30], expressing concerns about the EU's missed opportunity to regulate AI comprehensively by taking into consideration the principle of non-discrimination.

The evaluation of AI technologies is contested that relies on a solely *ex-post* evaluation of their risks which the statement judges incapable of grasping the true consequences and impact on human rights. Similarly, the statement suggests that a subjective approach to AI should be preferred as a means of: conferring to the proxy a central role in depicting the new form of discrimination; enlarging the list of suspected grounds of discrimination to not be limited to a selected and non-exhaustive enumeration of factors of discrimination. The statement also proposes the enlargement of prohibited AI technologies and the explicit safeguard of new rights associated with AI. This is the case with the right to explainability, which should go hand in hand with that of trustworthy AI, and the right to access *ad hoc* judiciary systems.

More generally, it seems that the EU Parliamentary Commissions and the NGOs' signatories of the statement wish to unleash the too-tight connection between AI and privacy, and vice versa, strengthening the latter

[30] The text can be read at the following link: https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf.

to the human rights realm, where AI has become as rapidly dangerous as the law demonstrated its failure to understand and tackle its prejudicial consequences.

More recently, 2022 saw the publication of the fourth compromised text and, on 25 November 2022, the EU Council adopted its approach to the AIA to enter into negotiation with the EU Parliament to finalize the procedure.

Worthy of mention is the new version of Article 6(3), which sets out the respect of fundamental rights as one of the main criteria to categorize AI technologies. The Article reads as follows: "AI systems referred to in Annex III shall be considered high-risk unless the output of the system is purely accessory in respect of the relevant action or decision to be taken and is not therefore likely to lead to a significant risk to the health, safety or fundamental rights".

Alongside this broader list of criteria to exclude the recurrence of high-risk AI technologies, the text provides a review of the list of high-risk AI systems and offers an updated definition of AI[31].

In particular, this last option is under debate as it significantly narrows down the definition of AI by way of limiting the scope of the text to machine learning systems, therefore placing a variety of other AI technologies outside the ambit of application of the proposed regulation.

Besides the legislative process, what matters is to place discrimination within the ongoing scenario. In other words, to verify whether and to what extent discrimination is truly considered in the text and rightly interpreted as one of AI's unwanted risks.

Not included in the original text, recent developments between November and December 2022 highlighted the EU's closer attention to non-discrimination, which came more into play compared to the previous and exclusive mention of the prohibition of profiling set out under Article 22 of the GDPR[32].

[31] Beyond the EU's debate, on the definition of AI see the literature on computer science, which however does not endorse a unanimous definition. Among others, see Aa.Vv, *Artificial Intelligence & Human Rights: Opportunities and Risks*, Berkman Klein Center for Internet & Society, Harvard University, 2018; S. Samoili, M. López-Cobo, E. Gómez, G. De Prato, F. Martínez-Plumed, B. Delipetrev, *Defining artificial intelligence*, European Commission, 2020; H. Surden, *Artificial Intelligence and Law: An Overview*, in *Georgia State University Law Review*, 2019, 1319 ff.; *High-Level Expert Group on Artificial Intelligence*, *A definition of AI: Main capabilities and scientific disciplines*, 2019, link: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf.; P. Boucher, *Artificial intelligence: How does it work, why does it matter, and what can we do about it?*, link: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf; within the constitutional law debate, see, among others, C. Casonato, *Intelligenza artificiale e giustizia: potenzialità e rischi*, in *DPCE Online*, 2020, 3369 ff.; F. Donati, *Intelligenza artificiale e giustizia*, 2020, 415 ff.; A. D'Aloia, *Il diritto verso "il mondo nuovo". Le sfide dell'Intelligenza Artificiale*, in *Rivista di BioDiritto*, 2019, 3 ff.

[32] Article 22 concerns *Automated individual decision-making, including profiling*, that is the only provision of the GDPR that tackles at least one type of discrimination deriving from AI systems.

There are some noteworthy aspects of the compromised text on this approved in November 2022. First, the explicit mention of AI's liability to give rise to "new forms of discriminatory impacts"[33], seems to suggest that the EU is slowly acknowledging the existing specifics of AI-based discrimination. The second derives from the new text of Article 10, whose letter *f)* contains an explicit reference to discrimination prohibited under EU law. Whether the provision will foster a link between the AIA and EU anti-discrimination law has not so far been proved, but it certainly shows an inclination to connect AI and discrimination, which was, conversely, absent in the previous text as well as in the GDPR.

Similarly, Article 64 establishes an unprecedented connection between access to justice and protection against AI-derived discrimination. The Article provides that: "National public authorities or bodies which supervise or enforce the respect of obligations under Union law protecting fundamental rights, *including the right to non-discrimination*, about the use of high-risk AI systems [...] shall have the power to request and access any documentation created or maintained under this Regulation when access to that documentation is necessary for the fulfillment of the competences under their mandate within the limits of their jurisdiction".

Thirdly, under point No. 3 of Annex IV, *Technical Documentation referred to in Article 11(1)*, the proposal hinges on the need to access "[d]etailed information about the monitoring, functioning, and control of the AI system" concerning "the foreseeable unintended outcomes and sources of risks to health and safety, fundamental rights and discrimination".

In February 2023, the EU Parliament later proposed a compromised amended version of the AIA which, once again, struggles with definitions and the categorization of the prohibited high-risk AI technologies. The EU voted this version and a new text has been recently published[34] with an expected adoption of the definitive text by the end of this year or at the beginning of 2024.

The latest version of the text approved by the EU Parliament in June 2023 seems to be even more in line with the trend emerged in late 2022 with an evident inclination to include non-discrimination seriously into the discussion. On this, worth considering are some of the amendments adopted, that will hopefully be part of the definitive text.

In few words, and waiting for the outcomes of the discussion among the three EU institutions involved in the approval of the proposal (the so-called "trilogy"), the compromised version shows a less reluctant approach to address the challenges posed to equality and non-discrimination by AI technologies. In the same direction goes the amendment presented by the EU Council and the European Parliament rapporteurs to include a chapter aimed at filling the gap on the lack of effective remedies to react to AI-based human rights violations.

[33] See § 37 of the Preamble.

[34] Here's the Link to the latest version of the text published in June 2023: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

Besides the explicit reference to the safeguard of the principle of non-discrimination in the new version of the Recital No. 9 of the Preamble[35], the new Recital No. 16(1) emphasizes the risks of discrimination posed by AI especially in cases of categorization of individuals along human qualities such as gender, gender identity, race, age, disability and, more broadly, all the factors of discrimination provided under Article 21 of the EU Charter of Fundamental Rights. Same *rationale* is shared by further amendments such as Nos. 53 and 75, that require the AIA to be in compliance with existing EU law, including EU anti-discrimination law, and explicitly recognize that existing EU and domestic laws should be considered already applicable to AI. While the previous versions of the AIA were silent on the relationships with EU principles and laws on equality and non-discrimination, the new amendments start filling the gaps and containing the criticisms towards the AIA's lack of awareness towards discriminations deriving from AI systems[36].

A similar approach also features the amended text of Article 10, whose actual wording goes beyond the mere statement about the "examination in view of possible bias" to explain the possible outcomes of biased AI systems and, among these, the risks of discriminations in violation of EU law.

Moreover, and of major significance is the new text of Article 4 of the regulation that includes under letter e) a reference to the principles of "diversity, non-discrimination and fairness", formally requiring AI systems to be "developed and used in a way that includes diverse actors and promotes equal access, gender equality, cultural diversity to avoid "discriminatory impacts and unfair biases [...]". This last provision is perhaps the most important one as it finally properly includes the principle of non-discrimination among the guiding principles governing AI systems under EU law[37].

Whether the approach will be effective is far from being solved and questions remain open as to the extent to which the regulation of AI will complement or integrate existing anti-discrimination laws.

However, it should be noted that the EU recently chose to focus on a different, but rather unimportant and interconnected aspect of AI and its negative consequences on fundamental rights and non-discrimination.

[35] The text reads as follow: "systems should make best efforts to respect general principles establishing a high-level framework that promotes a coherent human-centric approach to ethical and trustworthy AI in line with the Charter of Fundamental Rights of the European Union and the values on which the Union is founded, including the protection of fundamental rights, human agency and oversight, technical robustness and safety, privacy and data governance, transparency, non- discrimination and fairness and societal and environmental wellbeing".

[36] The text also expands beyond EU law. Interestingly, amendments No. 35 about Recital No. 13 specifies that the normative standards adopted to govern high-risk AI systems should also be in compliance with those enshrined under the European Green Deal, the Joint Declaration on Digital Rights of the Union and the Ethics Guidelines for Trustworthy Artificial Intelligence (AI) of the High-Level Expert Group on Artificial Intelligence.

[37] See, also, amendment No. 88, that introduces the new Recital No. 53 and that makes an explicit reference to the risks posed by AI to people with disabilities and amendment No. 228 with respect to the new text of Article 5, paragraph No. 1.

In September 2022, the EU Parliament presented a draft proposal of a *Directive on non-contractual civil liability rules to artificial intelligence*[38]. As well as the detailed analysis of its contents which goes beyond the scope of the Article, it is interesting to highlight the decision of EU institutions to provide individuals badly affected by AI technologies with a set of rules inspired by the aim of enforcing their right to access justice. This is more than a welcome step, that complements the other (dark) side of the relationships between AI and discrimination, represented by the many difficulties encountered in bringing cases before national and supranational Courts, especially when liability is hard to prove, as well as the uneasy identification of those (if "human") responsible for the unintended and possible discriminatory outcomes of AI technologies.

The two acts and their interconnection with EU anti-discrimination law will hopefully reveal the divergent nature of AI-based discrimination, finally offering the chance to confer an autonomous dignity to this new form of discrimination.

3. The Council of Europe: from the CAHAI, the CAI, and the First Proposed Framework Convention on AI

"To avoid unjustified bias, a provision on respect for equal treatment and non-discrimination should be included"[39]

Besides the EU, the most prominent role in linking AI to non-discrimination in Europe was certainly played by the Council of Europe[40]. Despite the first treaty on AI and human rights being so far away, more than any other international organization the Council of Europe demonstrated its willingness to investigate AI with a view to its human rights implications. Reference is chiefly made to the draft convention on artificial intelligence, human rights, democracy and the rule of law, the *AI Convention.*

The Committee of Ministers paved the way to the establishment of two *ad hoc* bodies, that operated subsequently one after the other with the

[38] The text can be read at the following link: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0496.

[39] Cfr. the *Explanatory Memorandum of the Recommendation for a Council Decision authorising the opening of negotiations on behalf of the European Union for a Council of Europe convention on artificial intelligence, human rights, democracy and the rule of law*. The text could be read at the following link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0414.

[40] See, first and foremost, the work of the *High-Level Expert Group on Artificial Intelligence*, which suggested an interesting definition of AI consisting in "[a] set of sciences, theories and techniques whose purpose is to reproduce by a machine the cognitive abilities of a human being". Reference should likewise be made to the *European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment*, of December 2018, link: https://www.europarl.europa.eu/cmsdata/196205/COUNCIL%20OF%20EUROPE%20-%20European%20Ethical%20Charter%20on%20the%20use%20of%20AI%20in%20judicial%20systems.pdf.

ultimate goal of laying down a general guideline of principles governing AI internationally: the CAHAI[41] and the CAI[42].

The CAHAI, established on 11 September 2019 until December 2021, aimed at examining "the feasibility and potential elements based on broad multi-stakeholder consultations, of a legal framework for the development, design, and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law". In its most prominent document, *Feasibility Study on a legal framework on AI Design, development and Application based on CoE Standards*, of December 2020, the CAHAI shows its tendency and willingness to take AI-based discrimination seriously, putting it at the forefront of the main risks of AI technologies. The document thus highlights that "[d]iscrimination, the advent of a surveillance society, the weakening of human agency, information distortion, electoral interference, digital exclusion, and a potentially harmful attention economy, are just some of the concrete concerns that are being expressed". The research is worth mentioning as it chooses to frame AI-derived discrimination within the dogmatic category of proxy discrimination, abandoning therefore the classical notions of anti-discrimination law.

Alongside and before the CAHAI, the Council of Europe adopted a series of additional soft law documents, ranging from the *European Ethical Charter on the use of artificial intelligence in judicial systems and their environment*[43] by the European Commission for the Efficiency of Justice (CEPEJ) in 2018 to the *Guidelines on facial recognition* of 2021 by the Consultative Committee of the Convention for the protection of individuals regarding the automatic processing of personal data, where there is an

[41] Ad hoc Committee on Artificial Intelligence (CAHAI), link: https://www.coe.int/en/web/artificial-intelligence/cahai. The first report can be read at the following link: *https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016809ed062* and it was released in September 2020. The CoE enabled the CAHAI especially to: «make substantive progress in the drafting of its feasibility study on a legal framework by November 2020, with a view to starting in January 2021 a reflection on the elements of a legal framework that would be the subject of a broad multi-stakeholder consultation; this legal framework could regulate the design, development and application of AI that have a significant impact on human rights, democracy and the rule of law. It could also consider the desirability of consolidating existing standards through an interpretation of the norms, principles and values already enacted in this area or developing new standards required for the digital age. Finally, it would lay the foundations on which a number of initiatives and instruments could be further developed in the different sectors of activity of the Council of Europe, which remains indispensable to comprehensively address the challenges posed by AI applications in the relevant fields of activity of the Council of Europe; propose, simultaneously, complementary measures to operationalise the above-mentioned legal framework: in particular, reference could be made to the prior human rights impact assessment procedure, the means of validation or certification of algorithms and AI systems or the training and organisation of certain professions involved in the application of AI tools».

[42] Committee on Artificial Intelligence, link: https://www.coe.int/en/web/artificial-intelligence/cai.

[43] Reference is made to the CEPEJ's *European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment*, see the text at the following link: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

explicit mention of the discriminatory risks associated with these types of AI technologies. Interestingly, discrimination is depicted as "unintended", emphasizing, even implicitly, two critical points: the lack of intentionality of the programmer(s), which is taken for granted and presumed without acknowledging the role of the humans in building the data-set; the distance from AI's discriminatory effects and human conduct, as if discrimination could always be regarded as a likely implication of AI technologies without investigating what might stand behind its discriminatory impacts.

Currently, the CoE's commitment to draft a general legal framework on artificial intelligence revolves around the activities of the newly established ad hoc Committee, the CAI.

Among its most prominent initiatives, in January 2023 the CAI presented the first international convention on artificial intelligence and human rights[44].

The "Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law"[45] is the result of the acknowledgment by the CoE of the twofold nature of AI, that of being simultaneously a tool to promote "human prosperity as well as individual and social well-being by enhancing progress and innovation", but, also, the cause of illegitimate interferences "with the exercise of human rights and fundamental freedoms, undermining democracy and violating the rule of law".

What is paramount here the most is that for the first time non-discrimination is recognized as one of the most relevant principles that should guide the use and implementation of AI technologies.

Under Article 3, the Draft Convention states that: "[t]he implementation of the provisions of this Convention […] shall be secured without discrimination on any ground such as sex, gender, sexual orientation, race, color, language, age, religion, political or any other opinion, national or social origin, association with a national minority, property, birth, state of health, disability or another status, or based on a combination of one or more of these grounds".

Not only does the Draft Convention choose to strengthen the link between AI and discrimination, but it also places the safeguard of non-discrimination among its General Principles, acknowledging the seriousness of AI's risks on the guarantee of equality among human beings and social groups.

Such a choice is remarkable in that it recognizes the centrality of non-discrimination in the broader AI discourse, finally conferring to it the place it deserves.

[44] At present, it should be considered that in July 2023, the CAI published an updated draft, the so-called *Consolidated Working Draft* that can be read at the following link: https://rm.coe.int/cai-2023-18-consolidated-working-draft-framework-convention/1680abde66.

[45] For an insight on the text, see the following link: https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.

Additionally, the CoE's approach towards AI goes hand in hand with the most recent initiatives of UNESCO[46] and the United Nations, whose 2022 Principles for the Ethical Use of Artificial Intelligence[47] shows a similarly strong commitment to building strategies to cope with AI's harm to the principle of equality and non-discrimination.

4. The Global Scale. The Judiciary on AI-Based Discrimination: Few Cases but Responses Worth-Mentioning

An alternative but complementary way to look at how AI intersects with discrimination is through the investigation of how cases are brought and judged before Courts. On this, there is one key element to consider, which deals with the scarce case law existing on AI and discrimination.

It is true that globally AI systems have rarely been challenged before Courts and even less so when the legal question surrounding the unintended discriminatory effects of AI. Nevertheless, the few cases decided by national Courts offer an insightful overview of the criticisms and difficulties of proving and, eventually, sanctioning AI technologies.

Before going into more detail on some of these judgments, it is useful to summarize the issues faced by Courts and by applicants in choosing to bring their case before the judiciary. These are the *ex-post* reconstructions of the discriminatory conduct; the identification of the person or entity responsible, together with the machine; the identification, once again, and selection of the proxy(ies) and its (their) correlation(s) with one or more traditional factors of discrimination; the proof of the discriminatory treatment and/or effects.

In as much as AI-based discrimination distances itself from traditional discrimination, the proof of its recurrence during trials turns out to be particularly complex. The few known cases are therefore of great relevance, in that they might contribute to building a pathway toward a set of guidelines for the understanding of the hidden implications of AI-based discrimination.

Moving on to the few cases decided by national Courts, noteworthy are common attempts to reconcile AI-based discrimination within anti-discrimination laws and the commitment to ascertain who should be considered liable for discrimination caused by AI.

Along with the well-known Compas case[48], another case was decided in Italy with the condemnation of the "Deliveroo" rider. In December 2019,

[46] Reference is made to UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, see the full text at the following link: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[47] The Guidelines were issued by the Inter-Agency Working Group on Artificial Intelligence. The text can be read at the following link: https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf.

[48] For a comment on the Compas case, see J. Angwin, J. Larson, S. Mattu, *Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*, in *ProPublica*, 2016; S. Carrer, *Se l'amicus curiae è un algoritmo: il chiacchierato caso*

the Tribunal of the city of Bologna held that the software used by Deliveroo to allow riders to access, reserve and cancel their work sessions was discriminatory[49]. According to the judge, the software treated all riders equally regardless of the reasons behind the cancellation of the work session. Therefore, the software did not differentiate among cases and the cancellation of work sessions, risking impairing fundamental rights when the motive of the conduct was related to the exercise of a right of constitutional relevance. In the "Deliveroo" case, the complaint focused on the lack of protection of the right to strike, guarantee of which was neglected by the "blindness" of the machine, unable to identify and distinguish between different requests for cancellation.

According to the Tribunal, the difference in treatment amounted to indirect discrimination, in that Deliveroo knew how the software worked, without deliberately wanting to discriminate. Such a construction of indirect discrimination is, nevertheless, peculiar. The Tribunal's decision hinges on the neutrality of the rule and on its disparate impact, but at the same time recalls a new and additional factor, and that is "Deliveroo"'s the knowledge of the likely discriminatory functioning of the AI system to the detriment of the workers. In other words, the Tribunal mixes objective and subjective elements that do not match with the ontological features of indirect discrimination, proving the specifics of AI-based discrimination.

Other interesting cases have been decided in recent years in Europe.

Loomis alla Corte Suprema del Wisconsin, in *Giurisprudenza Penale Web*, 2019, ff. An additional interesting case is represented by the United States Supreme Court's case *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc.*, with a comment by L. Rodrigez, *All data is not credit data: closing the gap between the fair housing act and algorithmic decisionmaking in the lending industry*, in *Columbia Law Review*, 2020, 1843 ff. Another interesting case was decided, this time, in Canada. The case is *Ewert c. Canada*, 13 June 2018, SCC 30, [2018] 2 S.C.R. 165. The case was about a detainee of indigenous origin who had challenged the discriminatory nature of the software *Correctional Service of Canada* (CSC) used for risk assessment purposes. The Canada's Supreme Court acknowledged, in particular, that: "[r]ecent reports indicate that the gap between Indigenous and non-Indigenous offenders has continued to widen on nearly every indicator of correctional performance. For example, relative to non-Indigenous offenders, Indigenous offenders are more likely to receive higher security classifications, to spend more time in segregation, to serve more of their sentence behind bars before first release, to be under-represented in community supervision populations, and to return to prison on revocation of parole: Canada [...]. It is thus clear that the concerns that motivated the incorporation of the principle set out in s. 4 (g) into the CCRA are no less relevant today than they were when the CCRA was enacted. In the face of ongoing disparities in correctional outcomes for Indigenous offenders, it is crucial, to ensure that the correctional system functions fairly and effectively", §§ 60, 61.

[49] Tribunal of Bologna, 31st December 2020 with the comment of D. Testa, *La discriminazione degli algoritmi: il caso Deliveroo, Trib. Bologna, 31 dicembre 2020*, in *IusinItinere.it*, 26 January 2021. On the potential discriminatory outcomes of app used by food delivery platform, see, also, the case of Uber discriminatory app against disable people in the case *National Federation of the Blind of California et al v. Uber Technologies Inc et al*, decided by the U.S. District Court, Northern District of California, No. 14-04086. Link: https://law.justia.com/cases/federal/district-courts/california/candce/3:2014cv04086/280572/37/.

Instead of sanctioning the difference in treatment and looking at the type of discrimination at stake, these judgments recognized the liability of the user for not having gathered enough information on how the AI system operated.

In the case delivered by the Hague District Court in March 2020[50], the judge held that the "SYRI" software (System Risicoindicatie), used to prevent and combat fraud in the interest of economic welfare, violated Article 8 of the European Convention of Human Rights. Despite the fact that the Court did not call into question the discriminatory implication of the SYRI legislation, it nevertheless unraveled the discriminatory potentials of AI systems used to define human behavior, stating that the software was "insufficiently transparent and verifiable".

Similarly, in a case concerning the facial recognition system "AFR Locate", the Court of Appeal of Great Britain and Wales[51] sanctioned South Wales police for having made use of software for public surveillance reasons without having previously disclosed its functioning. In this sense, the Court argued that South Wales police failed "to address the potential for gender and racial bias" and this was enough to prove its liability. In the Court's view, therefore, to prove the liability of the user it is merely required to demonstrate their lack of knowledge of the data fed to the machine. Following a reasoning that could also be replicated in further judgments, the Court emphasized that all users of AI systems should be fully aware of the mechanism and functioning of the machine as well as the data it bases its decisions on. The lack or partial knowledge of the AI system on the part of the user is enough to prove its liability besides their involvement in the creation of the machine.

Finally, in 2108 the Finnish National Non-discrimination and Equality Tribunal[52] challenged AI-based discrimination – classified as direct and multiple discrimination – that a man was the victim of who had been refused credit by a bank. The judgment is noteworthy as the Tribunal focused on the specifics of AI-based discrimination suggesting a parallelism with statistical discrimination considered more in line with that connected to AI.

Along with Europe, the United States is also witnessing a slow increase in judicial cases on AI-based discrimination. One interesting and pending case filed in February 2023, *Mobley v. Workday, Inc.*, concerns

[50] Reference is made to the case No. C/09/550982 /HAZA 18-388, 5th February 2020. The judgment might be consulted at the following link: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878. On this, see A. Rachovitsa, N. Johann, *The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case*, in *Human Rights Law Review*, 2022, 1 ff.

[51] Case *R (Bridges) -v- CC South Wales & ors*, 11th August 2020. Link: https://www.judiciary.uk/wp-content/uploads/2020/08/R-Bridges-v-CC-South-Wales-ors-Judgment.pdf.

[52] The case, decided by the *National Non-discrimination and Equality Tribunal of Finland*, No. 216 of 2017, 21st March 2018, can be read a the following link: https://www.yvtltk.fi/material/attachments/ytaltk/tapausselosteet/45LI2c6dD/YVTltk-tapausseloste-_21.3.2018-luotto-moniperusteinen_syrjinta-S-en_2.pdf.

alleged discrimination in hiring argued by the plaintiff, a disabled African American man of 40 years old who claimed to have been the victim of discrimination based on the protected ground by the software "Workday" used by the respondent company. The case is ongoing, but it will be interesting to monitor the outcome in the light of the hopefully acknowledged specificity of discrimination resulting from AI.

Lastly, it should be mentioned that there have been cases of AI-based discrimination, that did not go before the Courts. These examples show alternative ways to detect and counter AI-based discriminations, avoiding the difficulties in accessing the judiciary and the length of judicial proceedings. One case was the algorithm used in Washington DC to ask citizens to let the City Council know the metropolitan areas in need of renovation or infrastructure interventions, which made use of the Zip Code as a proxy. Since the software was available only on smartphones more commonly available in the central neighborhoods of the city, the software resulted in discrimination based on social class and race. Other cases to remember include the robot incapable of identifying Asian faces used in New Zealand or, once again in the United States, the Optum algorithm used to identify high-risk patients and proved discriminatory on racial grounds.

Conclusions: Misconceptions, Lessons Learned and Moving Forward: the "new" AI-Based Discrimination

Whether AI might impact human rights and the principle of non-discrimination is nowadays globally recognized.

Nevertheless, legislative responses are still far from acknowledging the specifics of discrimination deriving from AI. On their side, Courts are struggling between resorting to traditional anti-discrimination laws and shifting the issue, sanctioning human rights violations other than differences in treatment caused by AI technologies.

It is also true that an inversion of the trend has been recorded in recent times registered.

The Council of Europe presented its first International Convention on AI, choosing to isolate the principle of non-discrimination and showing a willingness to take AI-based discrimination seriously. Likewise, even not so explicitly, the European Union is slowly making space for a more robust human rights approach to AI also in the light of the principle of non-discrimination, which suggests that it could be advisable for a more fruitful exchange between the two organizations in their normative approaches to AI.

Despite the efforts, which are not limited to Europe, global trends are not entirely in accordance with the fight against AI discrimination. And the reason may be found in at least two misconceptions about the ways in which AI discriminates.

The first revolves around the diversity between AI-based and human-driven discrimination. The role of the machine in-between human action and the effects is obscured and often dismantles the causal link between the act and the effects. The different construction of the interaction between the action and the discriminatory effects, which is central to the proper understanding of the new type of discrimination, is instead generally neglected.

The second misconception is the ignored inadequacy of the classical categories and mechanisms of anti-discrimination law to counter AI-based discrimination. New forms of discrimination, such as proxy discrimination and unconscious difference treatment, call for alternative strategies which should rely almost solely on the effects rather than on the conduct. The said challenging reconstruction of the relationship between the human and the machine on an *ex-post* basis requires to depart from the scrutiny over the act to hinge instead on the effects.

Put differently, the more difficult the first is to trace, the more an effective strategy should look at the second and sanction the former only when the latter, the effects, prove to be discriminatory. In other words, the lack of intentionality, which often characterizes AI-based discrimination, should not impair the definition of an act as being discriminatory when directly discriminatory on a protected ground even if unintentional. This is an example of a form of direct AI discrimination, that diverges from the ideal type of disparate treatment, but which, nonetheless must be defined as such and sanctioned by the law.

One more note on this last point.

The above-mentioned deference to the effects, since they are capable of signaling the recurrence of AI-based discrimination, does not mean that the concept of indirect discrimination is always a suitable alternative to describe how AI discriminates. Quite the opposite. Neither the phenomenology of indirect discrimination is capable of grasping all the possible manifestations of AI-based discrimination. Proxy discrimination is exactly the symbolic representation of the failure of the dichotomy of direct/indirect discrimination to describe AI-based discrimination, as it simultaneously mixes both forms: it hinges on an element directly or indirectly linked to a protected ground; it is unintentional; it produces discriminatory effects following a likewise discriminatory conduct.

Proxy discrimination is thus placed at the crossroads between direct and indirect discrimination. The proximity with the concept of discrimination by association should eventually suggest that anti-discrimination law should abandon its original roots to explicitly connect AI-based discrimination with this latter form of discrimination[53].

Except for the reluctance of legislators and Courts to address AI-based discrimination moving from a realistic understanding of its traits, there are still some lessons learned worth considering.

First, the opportunity to integrate extra-legal knowledge in the study of how AI discriminates, incorporating research on statistics and economics. In fact, AI-based discrimination has a great deal in common with the concept of statistical discrimination, which is a type of discrimination that is insufficiently considered by legal studies, but quite similar to that caused by AI. From this angle, an exchange of knowledge, that relies on the concept of statistical discrimination, could provide a better way to redefining AI-

[53] On this, see S. Watcher, *Affinity Profiling and Discrimination by Association*, in *Berkeley Technology Law Journal*, 2020.

based discrimination according to a diverse categorization, leaving behind the inadequate categories of anti-discrimination law.

Secondly, with regard to accountability[54], the choice of some Courts to place the liability on the side of the user without proving their intentionality is certainly a remarkable one. On the contrary, Courts ask for the proof of the user's indifference toward the discriminatory effects and the lack of investigation into the functioning of the machine, which significantly simplifies the process and the identification of those responsible for AI-based discrimination.

Lastly, and conclusively, the experience demonstrates the heterogeneity of AI-based discrimination which, as noted, features traits that can no longer be understood within the constraints of anti-discrimination laws.

This global phenomenon calls for global responses that are nonetheless far from becoming reality in the near future.

Although legislators on a variety of levels are actively dealing with AI, they have not yet succeeded in untangling the exclusive link between AI and privacy to recognize that of AI and human rights.

Only the Council of Europe has rightly grasped the centrality of the risks AI poses to the principle of non-discrimination.

The hope is that the leading example of the Council of Europe will spread on a global basis and that not only will legislators favor the enactment of laws capable of acknowledging the specifics of AI-based discrimination but that they will also strengthen individuals' right to access justice so severely endangered in the face of a phenomenon that until now has not been adequately understood and contravened.

Costanza Nardocci
Dipartimento di Diritto Pubblico Italiano e Sovranazionale
Università degli studi di Milano
costanza.nardocci@unimi.it

[54] Accountability represents one of the major challenges in the current debate on how to prove AI-based discrimination before Courts. On this, see, among others, the study proposed by C. Wilson et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, 2021, https://evijit.github.io/docs/pymetrics_audit_FAccT.pdf.