Rivista di
Diritti Comparati

Addressing AI-driven gender discrimination: the role of the forthcoming EU AI Act and Corporate Social Responsibility[*]

Federica Fedorczyk

1. *Introduction: artificial intelligence and gender discrimination.*

The increased presence of Artificial Intelligence (AI) in our everyday lives is central to the transformation of our economy and society. AI creates new opportunities: it develops fast, and it is changing and improving many areas of our lives. However, some scholars raised serious concerns about the potential biases in dataset used by AI tools, which might result in altered predictions and connections. Indeed, such methods may be prejudicial in terms of gender, exacerbating inequalities and discrimination[1].

When we think about discrimination in AI, it rarely comes to mind to talk about gender discrimination: this side of discrimination is overlooked, but it still plays a crucial role in the new universe created by AI. Differently from race discrimination, gender discrimination arises not only from the collected data but also from the stage in which AI is designed, where female

[*] L'articolo è stato sottoposto, in conformità al regolamento della Rivista, a *double-blind peer review*.

[1] To have a complete overview on discrimination and artificial intelligence, see T. Wischmeyer - T. Rademacher (eds.), *Regulating Artificial Intelligence*, Cham, 2020; R. K. E. Bellamy et al., *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, in *IBM Journal of Research and Development*, 2018, 4/5; J. Burrell, *How the machine 'thinks': understanding opacity in machine learning algorithms*, in *Big Data & Society*, 2016, 1; T. Gebru et al., *Datasheets for Datasets. Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency*, in *Machine Learning*, 2018; R. Benjamin, *Assessing risk, automating racism*, in *Science*, 2019, p. 421; S. Hajian et al., *Algorithmic bias: from discrimination discovery to fairness-aware data mining*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining- Association for Computing Machinery*, San Francisco, 2016, p. 2125.

representation in the software engineering profession in global terms ranges from 3 to 7% (becoming 27% only in the United States)[2].

Researchers have noticed that human gender bias found its way to replicate itself into AI systems: in this regard, as it has been provocatively said, «AI is replicating the same conceptions of gender roles that are being removed from the real world»[3]. Indeed, AI systems are based on models that are abstract representations of complex realities where much information is not considered: «Models, despite their reputation for impartiality, reflect goals and ideology. (…) Our own values and desires influence our choices; from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics»[4].

The pre-existing biases in our society affect the way we interact and the data which are used to train machine learning system: therefore, when AI technologies are developed and trained using biased data, they allow our own biases to be confirmed and preserved[5]. As data become increasingly derived from human sources, it is more likely that AI will possess the ability to discriminate. Indeed, the biases and stereotypes found in human society, which are a result of historical or institutional discrimination, are also present in the data and will consequently perpetuate those same biases[6]. Moreover, when human bias joins forces with machine learning bias, the side effects are multiplied, and the discrimination grows exponentially.

Potentially, every element of data about humans can be affected by gender bias in AI: one of the main causes is that women are not being well

[2] World Economic Forum, *Global Gender Gap Report 2018*, 17 December 2018, available at https://www.weforum.org/reports/the-global-gender-gap-report-2018; see also I. Santoemma, *Dialogo con Marzia Vaccari - I.A., una prospettiva femminista*, in *Rivista Arel*, 2021, p. 169, where the Autor reported that according to «Evans Data Corporation, in 2019 there were 26.4 million software developers worldwide of which 4.2 million were in the US, of these only 27.5 per cent are women. Globally, the number of women employed in software development does not reach 3 per cent, for other statistical sources 8 per cent» (translated into English by the author).

[3] World Wide Web Foundation, *Policy brief W20 Argentina, Artificial Intelligence: open questions about gender inclusion*, 2018 available at http://webfoundation.org/docs/2018/06/AI-Gender.pdf.

[4] C. O'Neil, *Armi di distruzione matematica: come i Big Data aumentano la disuguaglianza e minacciano la democrazia*, Firenze - Milano, 2017, p. 33 (Italian edition of the original English version: C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy"*, New York, 2016).

[5] In this sense, see D.J. Fuchs, *The Dangers of Human-Like Bias in Machine-Learning algorithms*, in *Missouri S&T's Peer to Peer*, 2018, 1.

[6]C. Nardocci, *Artificial Intelligence-based Discrimination: Theoretical and Normative Responses. Perspectives from Europe*, in *DPCE Online*, 2023, p. 2370.

represented in the data.[7] In this regard, as Gina Neff accurately stated «women are not at the table when technologies and data systems are being built and designed […] they are less likely to be the subject of political news, they are less likely to be contacted as sources of information, they are less likely to be on Wikipedia, they are less likely to be in our dataset, they are less likely to be in medical data»[8]. Therefore, when scientist build technologies that scan for massive sources of information, data collection leaves out a huge amount of information about women: it generates unfairness and, above all, it creates results which will look like they are technologically neutral and transparent, while they are not.

Gender discrimination is also a problem of language since our language suffers from a patriarchal approach. If on one hand it is true that we cannot pretend from AI something that is still not present in the real word - for instance, if in the "real" world a male word to indicate housewives is not used, why should this word be used in the AI world? -, on the other hand, it is important to reflect on which role we would like AI to assume. Precisely because AI has more instruments and it is capable of processing thousands of data in just few seconds, it should become a vehicle through which correct the imperfections present in the "real" word, by first detecting them and second trying to eliminate them. This of course will be possible only if more inclusive data training set will be created. Otherwise, AI would play just a passive role, creating at a larger scale all the different forms of discriminations already present.

[7] R. Abrahams, *Alexa, does AI have gender?*, 2018, available at: https://www.research.ox.ac.uk/Article/2018-10-15-alexa-does-ai-have-gender.

[8] G. Neff, former Professor at the Oxford Internet Institute at the University of Oxford, now Executive Director of the Minderoo Centre for Technology & Democracy at the University of Cambridge, during one of her interviews available at https://www.oii.ox.ac.uk/news-events/events/oii-neff-lecture/. Moreover, see UNESCO, *Artificial intelligence and gender equality: key findings of UNESCO's Global Dialogue*, 2020, p. 44, available at https://unesdoc.unesco.org/ark:/48223/pf0000374174: «Women make up only 17% of the biographies on Wikipedia and 15% of its editors, and there is evidence of gender bias in the language of Wikipedia entries, 26% of subjects and sources in mainstream Internet news stories are women (Who Makes the News - http://whomakesthenews.org/gmmp-campaign), there are significant gender gaps across the stages of academic publishing, citation and comment (NIH - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7112170/), and the rather dismal figures go on, particularly looking at historical data. Women's access to and different use of ICT (phones, internet access, digital literacy, and so on) is another significant factor for the representativeness of data. In 2017 there were 250 million fewer women online (EQUALS - https://www.equals.org/post/2018/10/17/beyond-increasing-and-deepening-basic-access-to-ict-for-women)».

Furthermore, there is a significant gender disparity in the AI workforce, since gender diversity is still not balanced in STEM subjects at school or university. A labour market which fails to reflect a diverse population will inevitably exacerbate existing inequalities, since it will persist to show just one side of the coin. As stated by Simon de Beauvoir: «Representation of the world, like the world itself, is the work of the men; they describe it from their own point of view, which they confuse with the absolute truth»[9].

To invert this trend, diversification of the AI workforce will be crucial to design and realize technology which is unbiased. At the same time, there is an urgent need for research which analyses public policies and legislation related to AI which will impact on gender equality. In this regard, the first step should be to recognize the existence of a process of data discrimination that reinforce inequalities and oppression: one of the questions to be answered is therefore «how conscious of this issue citizens and public authorities who are purchasing, developing and using these systems are?»[10].

The first aim of this paper is therefore to consider the legal framework on AI through a gender lens in order to consider how existing and future measures can be used to tackle gender equality. The precondition for a similar analysis is to uncover gender inequalities which arise from the use of AI and to identify which are the relevant gaps in the state of the art that need to be addressed.

2. *The EU legal framework: just a starting point. Regulatory gap in AI standards and the need for an autonomous category*

Direct and indirect discrimination are prohibited in treaties and constitutions. When an act directly discriminates people on the basis of a protected characteristic (for instance sex or race) direct discrimination is performed; when instead an act is in theory neutral, but in practice is

[9] C. C. Perez uses this passage from the S. De Beauvoir's masterpiece The Second Sex as the preface to the introduction of her last work C. C. Perez, *Invisible Women: Data Bias in a World Designed for Men*, New York, 2019.

[10] P. Peña - J. Caron, *Decolonising AI: A transfeminist approach to data and social justice in Global Information Society Watch 2019: Artificial Intelligence: Human rights, social justice and development*, in *APC*, 2019, p. 30.

discriminating against people on the basis of protected characteristics, indirect discrimination occurs[11].

More specifically, EU law defined direct discrimination as a situation in which «one person is treated less favourably than another is, has been or would be treated in a comparable situation»[12]: in the context of algorithms, if any of their elements are not neutral towards a protected ground, direct discrimination is performed.

Instead, indirect discrimination refers to situations «where an apparently neutral provision, criterion or practice would put member of a protected category at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary»[13]. However, the implementation of the prohibition of indirect discrimination appears to be difficult since the concept of indirect discrimination results in rather open-ended standards. Indeed, it has to be proven that a decision apparently neutral affects a protected group creating a disproportionate effect and differentiating treatment ends up being non-discriminatory if there is an objective and reasonable reason, supported by a legitimate aim, to treat in a different manner similar situation[14].

It is important to highlight that, as indicated by the prevailing view among scholars, discrimination perpetrated by humans and the one arising from AI are fundamentally distinct: AI-derived discrimination merits recognition as an autonomous category, challenging the conventional distinction and dichotomous relationship between direct and indirect discrimination[15]. However, in the EU context, for a long time, the main

[11] The literature concerning algorithmic discrimination and the need for an ethical AI is boundless: most of the research on algorithmics and discrimination is focused on the U.S. context, but recently also the EU context has been deeply analysed. For instance, to cite just a few: C. V. Eubanks, *Automating inequality: how high-tech tools profile, police and punish the poor*, New York, 2018; O'Neil, *op. cit.*; E. Ellis - P. Watson, *EU Anti-Discrimination Law*, Oxford, 2012; S. Fredman*, Direct and Indirect Discrimination: is there still a divide?*, in H. Collins - T. Khaitan (eds.*), Foundations of Indirect Discrimination Law*, Oxford – Portland, 2018; S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, 2018.

[12] Art. 2(2)(a) Directive 2000/43/EC.

[13] Art. 2(2)(b) Directive 2000/43/EC.

[14] In this regard, Art. 2(2)(b) of the Racial Equality Directive 2000/43/EC prescribes that a practice will not constitute indirect discrimination if it is «is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary».

[15] For a comprehensive understanding of the rationale behind the distinct nature of AI-derived discrimination as an autonomous category, see the work of C. Nardocci, *Intelligenza artificiale e discriminazioni*, in *Rivista del Gruppo di Pisa*, 2021, p. 9-60.

instrument to protect people against AI-driven discrimination has been non-discrimination law and data protection law[16]. Still, the complexity and heterogeneity of AI-based discrimination necessitates lawmakers to adapt existing anti-discrimination laws in consideration of the specific characteristics of this new kind of discrimination[17]. Indeed, for different reasons the mere ban on the use of protected characteristics is not always sufficient to prevent AI-driven discrimination[18].

First, often it is not possible to precisely know the types of data used by the software: this opacity is commonly known with the term "black box". This term has a dual meaning: the more traditional one, namely the recording device used to track movements of cars, trains, and planes; and the "new" one, the one that interests us the most: black box evokes a system whose workings are mysterious since we can observe its inputs and outputs, but we cannot tell how one becomes the other[19]. Indeed, the complexity of

[16] The references are to the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1 and to the various directives that address discrimination: Directive 2000/43/EC against discrimination on grounds of race and ethnic origin; Directive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation; Directive 2006/54/EC equal treatment for men and women in matters of employment and occupation; Directive 2004/113/EC equal treatment for men and women in the access to and supply of goods and services; Directive Proposal (COM(2008)462) against discrimination based on age, disability, sexual orientation and religion or belief beyond the workplace.

[17] C. Nardocci, *op. cit.*, p. 2372.

[18] In this regard see M. Coeckelbergh, *Ethics of artificial intelligence: Some ethical issues and regulatory challenges*, in *Technology and Regulation*, 2019, p. 31. The Author contended that addressing bias presents a complex challenge. It remains unclear what precise measures should be taken to minimize bias and who should be responsible for implementing them. In cases where existing regulations are perceived as inadequate, there is a need to justify the necessity of new regulations. For instance, in the realm of data protection and privacy, as well as in the context of transparency and explainability, divergent perspectives emerge. Some argue that the General Data Protection Regulation (GDPR) offers comprehensive and enforceable legislation, Conversely, others assert that the GDPR lacks sufficient safeguards against the risks associated with automated decision-making, especially concerning explainability, since under the GDPR, individuals have a right to information, but it does not mandate complete explainability. For a deep analysis of this last point, see also 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, p. 99.

[19] F. Pasquale, *The Black Box Society. The Secret Algorithms That Control Money and Information*, Cambridge, 2015

the coding normally prevents transparency and accessibility and even when the coding itself could be transparent, often the products are protected by intellectual property law, so both developers and clients are generally uninclined to open them up to the public gaze[20].

Second, even imagining a situation in which the ban on the use of protected characteristics is complied with, the risk of discrimination is still not neutralized. Indeed, even if one does not directly use the prohibited characteristics since they are considered discriminatory, one can use other categories correlated with them, which somehow manage to circumvent the prohibition on the use of the protected categories. The reference is to proxy discrimination. In the context of AI discrimination, a proxy refers to a variable or feature that is used as an indirect or substitute measure for another characteristic that may be sensitive or protected. As stated by Prince and Schwarcz, «proxy discrimination is a particularly pernicious subset of disparate impact» that consists of a seemingly innocuous practice that however causes a disproportionate damage to individuals belonging to a protected group[21]. The issue stems from the inclusion of what are known as "redundant encodings" in the datasets. These redundancies refer to instances where membership in the protected category is encoded in other data points, and these encoded aspects ultimately correlate with the same protected category[22].

Just to cite one example, machine learning is increasingly being used by insurance companies to determine risk factors for car accidents. In the past, groups such as men and women were used as factors for measuring risk, due to the complexity and cost of gathering more granular information. However, this approach raises legal and ethical concerns, as it could be considered discriminatory and in violation of fundamental principles of equality.

Indeed, equality between men and women is a fundamental principle of the European Union: Articles 21 and 23 of the Charter of Fundamental Rights of the European Union prohibit any discrimination on grounds of sex and require equality between men and women to be ensured in all areas. Therefore, gender should not be used as a relevant factor. In this regard,

[20] D. M. Taramundi, *Discrimination by Machine-Based Decisions: Inputs and Limits of Anti-Discrimination Law*, in B. Custers - E. Fosch-Villaronga (eds.), *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, New York, 2022, p. 76.

[21] D. Schwarcz - A. E. R. Prince, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, in *Iowa Law Review*, 2020, p. 1260.

[22] S. Barocas - A. D. Selbst, *Big data disparate impact*, in *California Law Review*, 2016, p. 691.

the Court of Justice of the European Union (CJEU) in 2011 has even explicitly found that «The use of actuarial factors related to sex is widespread in the provision of insurance and other related financial services. To ensure equal treatment between men and women, the use of sex as an actuarial factor should not result in differences in individuals' premiums and benefits» (CJEU, 1 March 2011, C-236/09, *Association Belge des Consommateurs Test-Achats* and Others).

Still, while gender in the EU is considered a protected characteristic, and therefore cannot generally be used as a relevant factor, there is the risk that machine learning uses other indirect proxies (such as tastes and behaviours) that are correlated with gender to estimate risks. This would be the case for example if tastes for certain types of sports or cars that are held prevailingly by men are used to estimate given risks in the insurance sector[23]. In this way, the ban on the use of protected characteristics is circumvented and discrimination is still performed. Moreover, frequently, the correlation between the protected characteristic and the proxy used by the AI is ambiguous or, worse, the proxy itself is unknown[24]. Furthermore, the resultant prejudice is typically an unintended consequence of the algorithm's implementation rather than a deliberate decision made by its programmers, and, as a result, it can be particularly challenging to pinpoint the origin of the problem or provide a clear explanation[25].

Algorithms can discriminate even when they are not instructed to discriminate. In this regard, when an algorithm produces discriminatory outcomes regardless of the intention to discriminate and thus the effects are discriminatory, indirect discrimination occurs.

For instance, in the first Italian judicial decision on algorithmic discrimination, the Court of Bologna presumed the existence of indirect discrimination performed by the algorithm of Deliveroo platform in managing the delivery riders' workflows until November 2nd, 2020[26].

More specifically, the Court stated that by applying an apparently neutral provision - the contractual regulation on the early cancellation of

[23] R. Xenidis - L. Senden, *EU non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination*, in U. Bernitz et al. (eds.), *General Principles of EU law and the EU Digital Order*, The Netherlands, 2020, p. 5.

[24] C. Nardocci, *op. cit.*, p. 2373.

[25] S. Barocas - A. D. Selbst, *op. cit.*, p. 691.

[26] To have an overview on the case with a precise analysis of the possibility of framing algorithmic discrimination as a case of direct or indirect discrimination see M. Barbera, *Discriminazioni algoritmiche e forme di discriminazione*, in *Labour & Law Issues*, 2021, p. 3 ss.

booked sessions - a certain category of workers (and specifically those participating in trade union abstention initiatives) was put in a position of potential particular disadvantage[27].

Furthermore, it goes without saying that since machine learning algorithms have the ability to acquire their own knowledge by extracting patterns from data, the risk of indirect discrimination is even greater since there is the additional risk of reproducing existing patterns of inequality in ways unintended by their designers[28].

More specifically, there are several ways in which bias can pervade algorithmic design during the developing process: for instance, the software can reflect biases that are held by its developers[29]. Imagine an algorithm developed to find out who is the best candidate: the output would certainly depend on the definition given to the meaning of "best" and the output will vary according to it.

Prejudices can also influence the labelling process: if a white man wearing a white overall is likely to be identified as a doctor, women wearing the same dress will be more likely categorized as nurses and the externalization of similar data labelling might lead to prejudiced outcomes on a bigger scale[30].

Therefore, it is clear that algorithms cannot be considered neutral and that it is extremely difficult most of the time to detect the discrimination they perform. Risks of discrimination are indeed concealed both in the choice of the outcome that is entrusted to the algorithm and in the selection of the input information used to train it (candidate predictors), as well as in the training procedure used (including the training procedure used).

The promise of unbiased decisions remains thus just a promise and the mere fact that the decision is automated does not ensure a higher degree of objectivity.

Recently, EU provided more specific regulation on discrimination related to artificial intelligence, machine learning and robotics. In this regard, the European Parliament provided European guidelines on ethics in

[27] M. Barbera, *op. cit.*, p. 11.

[28] S. Barocas - A. D. Selbst, *op. loc. ult. cit.*

[29] S. U. Noble, *op. cit.*.

[30] On the consequences of similar stereotypes see M. Kay et al., *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations,* in Association for Computing Machinery, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing System*, New York, 2015.

artificial intelligence (2019)[31], with the aim of building a "human-centric" approach to AI, that is respectful of European values and principles. Furthermore, on 19 February 2020, the European Commission presented the "White Paper on Artificial Intelligence: a European approach to excellence and trust", which sets out some policy objectives: for instance, to launch an EU-wide debate on the use of remote biometric identification and to require high-risk AI systems to be transparent, traceable and under human control[32].

Nevertheless, the legal framework on gender equality in the use of AI appeared far to be complete: indeed, the UNESCO report on Artificial Intelligence and Gender Equality (2020) showed that AI normative instruments or principles that successfully address gender equality are either inexistent or insufficient[33]. If fairness, countability and transparency are explicitly named, gender equality is often only implicit and gender bias are not directly address.

Gender bias, according to social psychologists[34], may have some characteristics: it can be categorized into behavioural bias (discrimination), cognitive bias (stereotypes) and emotional bias (prejudice) and it consists of the preference for or prejudice against one gender over another. The preference can be conscious or unconscious and can also affect the whole society on a structural level.

However, frequently, gender-related issues are classified under the "fairness" category, even though it is well known that fairness has more than twenty different definition and it can include many different types of discrimination[35]. Therefore, although there are some explicit references to gender equality[36], generally it is reduced into more general goals, such as inclusiveness, social good, human well-being, which just implicitly include

[31] European Commission, *Ethics Guidelines for Trustworthy Artificial Intelligence*, 2019, available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[32] European Commission, *White paper on artificial intelligence a European approach to excellence and trust*, 2020, available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[33] UNESCO, *op. cit.*, p. 12.

[34] J. F. Dovidio - S. L Gaertner, *Intergroup bias* in S. T. Fiske - D. T. Gilbert - G. Lindzey (eds.), *Handbook of social psychology*, New York, 2010, p. 1084.

[35] C. Collett and S. Dillon, *AI and Gender: Four Proposals for Future Research*, Cambridge, 2019.

[36] A list of the normative texts which have explicit references to gender equality is provided by UNESCO, *op. cit.*, Annex 1.

women and realization of gender equality. For example, in the EU Ethics Guidelines aforementioned, the High-level expert group on artificial intelligence grouped "diversity, non-discrimination and fairness" in the same category, including in it the avoidance of unfair bias, the accessibility and universal design, and the stakeholder participation[37].

An exception in this respect is represented by the Union Network International (UNI) Global Union, which, in its 10 principles for ethical artificial intelligence, prescribes that «In the design and maintenance of AI, it is vital that the system is controlled for negative or harmful human-bias, and that any bias—be it gender, race, sexual orientation, age, etc.—is identified and is not propagated by the system»[38].

Along this line, something has recently changed also at a more institutional level, since the European Commission's Advisory Committee on Equal Opportunities for Women and Men issued an "Opinion on AI, opportunities and challenges for gender equality"[39]. The Opinion tried to explain first how AI risks to perpetuate gender inequalities and discrimination, and second how to mitigate these risks with policy actions. Furthermore, it tried to explore how AI can contribute to reduce gender inequalities.

In particular, the Opinion detected four main areas in which AI has contributed to generate gender discrimination: education, STEM sector, recruitment and data.

In education, AI can aggravate gender stereotypes since it can develop algorithms that mirror existing stereotypes, creating in this way discriminatory practices. Therefore, EC recommends Member States to encourage women to pursue a STEM education, raising awareness on career opportunities and highlighting role models[40].

[37] European Commission, *Ethics Guidelines for Trustworthy Artificial Intelligence*, 2019, p. 14.

[38] UNI Global Union, *Top 10 Principles for Ethical Artificial Intelligence*, available at: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.

[39] Advisory Committee on Equal Opportunities for Women and Men, European Commission, *Opinion on Artificial Intelligence - opportunities and challenges for gender equality*, 18 March 2020, available at https://ec.europa.eu/info/sites/info/files/aid_development_cooperation_fundamental_rights/opinion_artificial_intelligence_gender_equality_2020_en.pdf.

[40] Advisory Committee on Equal Opportunities for Women and Men, *op. cit.*, p. 7.

In STEM sector, the absence of diversity within AI development and data science teams is documented[41]. The majority of computer scientists and engineers in these fields are typically white males from Western countries: their individual viewpoints, and, at times, biases could influence the process and negatively affect women[42]. Furthermore, the male predominance might inhibit women's access to the area and make it more difficult for them to stay in these sectors.

Indeed, in an environment composed primarily by men it is more likely the creation of dynamics which accept vexations or push women to abandon their career[43]. Thus, the EC recommends taking positive actions in order to attract women to STEM careers and to ensure working conditions which can make it possible to conciliate work and family life[44].

Furthermore, it is of primary importance to invite Member States and stakeholders to take measures to prevent sexism: in this regard, a first important step could be the introduction of serious and severe consequences against perpetrators[45]. Indeed, it is well-established that workplace sexual harassment contributes to increased absenteeism and turnover rates, while also diminishing overall workplace productivity and job satisfaction[46]. Despite these known detrimental effects, it continues to persist at alarming levels and often goes unreported. Promising solutions to tackle this problem appear to revolve around the implementation of robust workplace policies that expressly prohibit sexual harassment, comprehensive training programs, and a transparent complaints process

[41] See, for instance, among other, C. Botella et al., *Gender diversity in STEM disciplines: A multiple factor problem*, in *Entropy*, 2019, p. 30; B. J. Casad et al., *Gender inequality in academia: Problems and solutions for women faculty,* in *STEM in Journal of neuroscience research*, 2021, p. 13.

[42] M. Coeckelbergh, *AI Ethics*, Cambridge 2020, p. 84.

[43] In this regard, the Advisory Committee on Equal Opportunities for Women and Men, *op. cit.* reported that 56% of female technical staff in the tech industry quit their career at the mid- level point, twice the resignation rate of men.

[44] In this sense, a recent good practice can be found in the Grande École du Numérique (Paris). As stated in GEN, "Favoriser la mixité dans le secteur du numérique" (February 2017), the School aims to promote gender equality in the digital sector and ensure women have access to opportunities on offer within the field. Accredited courses are therefore tasked with ensuring at least 30% of their student intake are female. The Grande École du Numérique also promotes courses that enable mothers to enroll thanks to family-friendly timetables.

[45] Advisory Committee on Equal Opportunities for Women and Men, *op. cit.*, p. 9.

[46] European Commission, Directorate-General for Employment, Social Affairs and Inclusion, Sexual harassment at the workplace in the European Union, Publications Office, 1999. See also C. R. Willness et al., *A meta-analysis of the antecedents and consequences of workplace sexual harassment*, in *Personnel Psychology*, 60(1), 2007, p. 127-162.

that shields employees from retaliation[47]. These measures collectively hold the greatest potential for reducing the prevalence of sexist behaviours in the workplace.

As far as recruitment is concerned, it is well known that companies use AI automated procedures to select candidates and many of them use a scoring system, based on specific requirements that are associated to positive evaluations. Risks for gender equality are connected mainly to historic bias and discrimination already existing in the real world. Moreover, due to the black box-issue, it is always difficult to discover motivations behind the hiring final decision. For these reasons, the EC recommends ensuring transparency in the HR sector concerning the criteria used in the recruitment process, since it is essential to guarantee the accessibility and at the same time to respect constitutional values, such as privacy and intellectual property[48].

The last focus is on the data, which are the primary means by which contrast gender inequality. Indeed, data must represent social diversity since high-quality data are the only possible basis for building an AI free from bias.

It is important to notice that self-learning algorithms work with correlation and often cause indirect discrimination. Therefore, the EC recommends using data disaggregated by gender and sex, in order to detect both direct and indirect discrimination against woman[49]. Moreover, Member States and stakeholders should invest in researching possibilities of non-discrimination by design. That means that regulation does not only concern the use of the AI systems, but also, and especially, the development of the technology, that should be informed the aforementioned principles from the from the outset and required their traceability at all stages[50].

In this regard, a lack of discriminatory intent is not sufficient to avoid discriminatory design: an active anti-discriminatory prospective should be adopted in the early stage of design to ensure non-discriminatory outcomes[51].

All the recommendations and steps required by the European Commission should be taken with a special attention to the monitoring

[47] J. Hersch, *Sexual Harassment in the Workplace*, IZA World of Labor, 2015.

[48] Advisory Committee on Equal Opportunities for Women and Men, *op. cit.*,p. 10.

[49] Advisory Committee on Equal Opportunities for Women and Men, *op. cit.*, p. 11.

[50] V. Dignum et al., *Ethics by Design: Necessity or Curse?*, in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, New York, 2018.

[51] To have an overview about discrimination by design see D.E. Wittkower, *Principles of anti-discriminatory design*, in *Philosophy Faculty Publications*, 2016.

phase: discrimination can affect all stages of the lifetime cycle of AI system and to detect them it is fundamental to monitor all stages (training, coding, new input data)[52].

What emerges from the analysis of the Opinion and the existing sources is the need to create a new and autonomous category dedicated to gender equality in the creation and the use of artificial intelligence. Gender equality should become not just a shade of other categories, but a primary and autonomous goal to be rapidly achieved, bearing in mind that AI can become an instrument to reveal discrimination and contrast them, instead of creating and amplifying them.

Indeed, a specific analysis on how policy and legislation should facilitate AI to work for gender equality is needed, especially to imagine an effective mutual understanding between technicians and policymakers regarding definitions of gender to build a common semantic. In other words, to tackle the main issues in an accurate manner, collaboration between policymakers, experts and technologists is crucial: ethical principles are not anymore adequate to ensure the implementation of new policies and legislations and therefore they should be accompanied by specific and context related standard directly applicable. The construction of an inclusive AI should be designed in order to ensure a dual objective, to use Villani's worlds: «First, to ensure that the development of AI technology does not cause an increase in social and economic inequality. Second to call on AI in order to reduce this. Rather than jeopardizing our individual trajectories and solidarity systems, AI must first and foremost help us to promote our fundamental rights, improve social cohesion, and strengthen solidarity»[53].

3. *The impact of the forthcoming EU AI Act*

In April 2021 the EU Commission tabled a "Proposal for a regulation laying down harmonized rules on artificial intelligence"[54] with the specific

[52] Advisory Committee on Equal Opportunities for Women and Men, *op. cit.*, p. 12.

[53] In March 2018, France presented its vision and strategy on AI. The French AI strategy is entitled *AI for humanity* and has been developed on the basis of the AI policy report, prepared by Cédric Villani, French deputy in the National Assembly. The reference is to the so-called "Villani Plan", available at https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

[54] European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, COM/2021/206 final.

object, among the others, to ensure that AI systems placed and used on the Union market are safe and respectful of existing law on fundamental rights[55].

The legal instrument of the regulation was chosen because uniform and harmonized rules for AI were needed: it was necessary to structure a market favourable to innovation, but at the same time guaranteeing citizens' rights. There was also the willingness to affirm the digital sovereignty of the European Union: that is why the Regulation was used and no room was left to Member States through the instrument of Directive. The Regulation, indeed, applies to all providers of services and goods that are sold and offered on the European market, regardless of where the providers are located.

Furthermore, the choice to use the normative tool of the Regulation to regulate AI systems in the EU was influenced by the Regulation's ability to have direct horizontal effects: EU Regulations that are sufficiently clear, precise, and relevant to the individual's situation, do not only impose obligations on EU Member States but also grants rights to individuals. Consequently, individuals can directly invoke EU law in both national and European courts, even in cases where there is no judicial remedy under national law. More specifically, the horizontal direct effect allows individuals to use EU law not only towards their State, but also towards other individuals. For this reason, the AI Act has been defined as «leap forward for horizontal artificial intelligence regulation»[56].

As outlined in the explanatory Memorandum, the horizontal nature of the proposal necessitates strict alignment with the current Union legislation: it is important to note that the Proposal guarantees alignment with the EU Charter of Fundamental Rights and the prevailing secondary Union legislation, which includes regulations on gender equality and non-discrimination[57]. Additionally, the proposal enhances existing Union regulations related to non-discrimination by introducing specific provisions aimed at reducing the potential for algorithmic bias[58].

However, at a closer look, in the original Proposal there was no further trace of gender equality, or rather, the word "gender" was only

[55] I have analyzed the aims, the structure and the evolving modifications of the AI Act in the following contribution: F. Fedorczyk, *AI legislation in flux: tracking evolving modification of the AI Act*, in *Diritti Comparati Blog*, September 2023.

[56] F. Lütz, *Gender Equality and Artificial Intelligence in Europe. Addressing Direct and Indirect Impacts of Algorithms on Gender-Based Discrimination*, in *ERA Forum*, 2022, 1.

[57] See Explanatory Memorandum of the Proposal, 1.2.

[58] See Explanatory Memorandum of the Proposal, 1.2.

mentioned once, in aforementioned explanatory Memorandum, where it was stated that the consistency of the Act is «also ensured with the EU Charter of Fundamental Rights and the existing secondary Union legislation on data protection, consumer protection, non-discrimination and gender equality»[59].

Similarly, the word "woman" was used only two times: first, in the part of the Memorandum that addressed fundamental rights and recall the need for the Proposal to enhance and promote the rights enshrined in the EU Charter of Fundamental Rights, among which it is cited the equality between women and men provided by Art. 23[60]. Second, in the Recital (36), where it was claimed that an area in which the use of AI systems need special consideration is the area of employment, workers management and access to self-employment, where the recruitment process performed with the contribution of AI system may perpetuate «historical patterns of discrimination, for example against women»[61].

At a first sight, this was certainly a missed opportunity to give space to issues related to gender discrimination linked to the use of AI systems. Indeed, the fact that the first draft of the Commission missed to address directly gender equality appeared to be a serious deficiency, which mirrored the EU's inability to deal with the issue in depth. If the public sector does not decide to tackle this issue, the risk is that the private sector will create a non-homogeneous discipline, composed by non-preceptive prescriptions and standards not always respected.

However, after the Commission adopted the Proposal on 21 April 2021 and the Council unanimously adopted its General Approach on 6 December 2022, in May 2023, the European Parliament introduced some amendments, which were adopted with a substantive majority vote on 14 June 2023, giving start to the Trilogue negotiations. The Parliament's version contains different changes, among which a new attention to gender issues.

More specifically, the term "gender" or "gender equality" has been introduced in different new Recitals. First, in the provisions concerning the prohibition of AI systems that categorise natural person by assigning them to «specific categories, according to known or inferred sensitive or protected characteristics are particularly intrusive, violate human dignity and

[59] European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, COM/2021/206 final, p. 4

[60] See Explanatory Memorandum, 3.5.

[61] Recital (36).

hold great risk of discrimination»[62]. In this respect, explicit references have been made to Art. 21 of the EU Charter of Fundamental Rights, and to Art. 9 of Regulation (EU) 2016/769, highlighting that the sensitive or protected characteristics mentioned include, among the others, gender and gender identity.

Second, in the new Recital 28 (Amendment 56) a specific reference has been made to the right of gender equality, since AI systems classified as high risk can cause a considerable negative impact on fundamental rights protected by the Charter, among which, the right to gender equality.

Also in the Amendment 67, that aims to modify the Recital 37 concerning high risk AI systems used to evaluate the credit score of natural persons in the access to and enjoyment of certain essential private and public services, the word "gender" is included.

Furthermore, some changes have been made also in the Articles of the Regulation. The Parliament introduced a new Art. 4 a) entitled "General principle applicable to all AI systems" in which it is stated that all operators falling under the Regulation shall make their best efforts to develop and use AI systems and foundation models in accordance with some core principles of the Union, among which «diversity, non-discrimination and fairness», and that AI systems should be developed and used «in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity»[63].

Another important provision is the new Art. 4 b) about "AI literacy"[64]. It states that the Union and the Member State shall act to ensure the development of a sufficient level of AI literacy, also by ensuring proper gender balance[65]. Similarly, the new Art. 57 a), in regulating the composition of the management board of the new AI Office, provides that the appointment of members and substitute members shall take into account the need to gender balance. These new provisions highlight the need to address the gender gap and the lack of diversity in the AI field. The

[62] Amendment 38, Recital 16a (new).

[63] See Art. 4 a (new).

[64] See Art. 4 b (new).

[65] In this respect, it is worth mentioning also the new Recital 9b concerning "AI literacy" that states: «it is therefore necessary that the Commission, the Member States as well as providers and users of AI systems, in cooperation with all relevant stakeholders, promote the development of a sufficient level of AI literacy, in all sectors of society, for people of all ages, including women and girls, and that progress in that regard is closely followed».

importance of gender balance lies in it being a prerequisite for diversity in opinions and the drafting of guidelines[66].

Finally, the Amendment proposed for Art. 69 explicitly states that Codes of conduct shall gave considerable consideration about the way in which the use of AI systems may have an impact or can increase diversity, gender balance and equality[67].

These new provisions mirror the willingness to incorporate a gender equality perspective into the AI Act and can be considered a step forward from the initial EU Commission's version. At the same time, they also prove that significant work still needs to be done to include women in this field and that more technical and sector specific regulations should still be created.

In December 2023, the forthcoming EU AI Act reached a new and almost final version following trilogue negotiations. However, the text was not made public until January 22, 2024, when an unofficial draft was published.

A first analysis of the text does not indicate any steps forward, but rather shows that steps backward have been taken, particularly concerning gender balance. The provision emphasizing the importance of gender balance in the new AI Office's board composition seems to be disappeared, as well as the reference to gender balance in Article 4b) on AI literacy. The term "gender balance" can now be found only in Recital 81 on codes of conduct, with no explicit reference in the main body of the Regulation, except in the related Article 69 on codes of conduct where the reference is on gender equality.

While this text is not final, and the official presentation by the EU institutions is still pending, it seems that there is still an urgent need, primarily at the legislative level, for provisions that establish precise requirements to ensure the respect of gender equality in the context of addressing algorithmic discrimination[68] and to ensure gender balance in AI governance.

[66] In this sense, see also the Opinion of the European Committee of the Regions, *European approach to artificial intelligence Artificial Intelligence Act (revised opinion)*, COR 2021/02682, OJ C 97, 28.2.2022, p. 60-85.

[67] Amendment 634 to Art. 69.

[68] F. Lütz, *op. cit.*, p. 47.

4. *Why a Regulation is needed: the existing gender-based discriminations*

There are already many examples which reveal that gender discrimination in AI is nowadays a real issue and therefore needs an urgent answer.

Notably, gender bias exhibited by AI systems is prominently observed across various components of Natural Language Processing (NLP) and the present analysis will primarily focus on describing instances of gender-based AI discrimination within the realm of NLP[69].

The intricacies of NLP systems, spanning the training data, linguistic resources, pretrained models (such as word embeddings), and underlying algorithms, can be susceptible to harboring gender biases[70]. NLP systems that exhibit bias in any of these components have the potential to generate predictions that are skewed towards a particular gender, and in some cases, may even exacerbate biases that already exist in the training data. Therefore, the growing dissemination of gender bias in NLP systems presents a risk of perpetuating harmful stereotypes in subsequent applications that have tangible repercussions in the real world.

Starting with one of the simplest examples, it can be easily demonstrate just using Google Translate: the translation from English to Turkish language (which is a gender-neutral language)[71] and back to English of same expressions as “She is the president” or “He is cooking”, turned out to be “He is the president” and “She is cooking”. In the first translation from English to Turkish, the gender pronouns were lost because the Turkish language is gender-neutral; but in the second translation, where Google Translate had the task of assigning pronouns to the jobs, the automatic result was inverted, with clear gender bias[72]. Even if recently this

[69] To have an overview on gender bias and Natural Language Processing see T. Sun et al., *Mitigating Gender Bias in Natural Language Processing: Literature Review, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, available at https://aclanthology.org/P19-1159/.

[70] J. Zhao et al., *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, 2018.

[71] A genderless language is a natural or constructed language that has no distinctions of grammatical gender. See Y. Suleiman, *Language and Society in the Middle East and North Africa*, Abingdon, 1999.

[72] The present experiment was reported in November 2017 by Quartz Magazine and analysed by G. Wellner - T. Rothman, *Feminist AI: Can We Expect Our AI Systems to Become Feminist?*, in *Philosophy & TechnologyPhilosy*, 2020, p. 191. On the same topic see also

bias has been corrected, other biases have even more recently emerged and reported by the social media Twitter in March 2021, regarding other languages characterized by neutral pronouns, such as Finnish and Hungarian. As a result of the media hype surrounding the re-sharing of this news, these latter biases have also been corrected. These examples are interesting because they demonstrate how human intent can drive AI to shift from being a tool for discrimination to one that respects gender equality[73].

To better understand how these kinds of bias appeared, it is necessary to talk about "word embeddings". Word embeddings are a list of numbers that encode information about words meaning and usage: algorithms learn these values by being given a range of training data composed by text, where words are used in their natural context. Normally, it is possible to compare these values in order to determine how are related two or more terms: however, if language models are trained on data that makes assumptions about the roles of men and women – amplifying gender biased perception – they can start to get wrong ideas and produce translations that project gender even when the original language does not specify one[74].

In order to further support this thesis, it is possible to take a simple test on a word embedding tool[75], which is able to find out solutions related to words: for instance, putting the inputs "Berlin" is to "Germany" as "Paris" is to "?", it will find out that the solution is "France".

At the same time, if we put "woman" is to "housewives" as "men" is to "?", the result will be "schoolteacher".

Similarly, showing the sexism behind the machine, when giving inputs "slut", the male result will be "douchebag", which notably means just "unpleasant person"[76].

Another experiment was made in 2013 through the communication campaign to raise awareness and combat sexist stereotypes performed by

M. Prates et al., *Assessing gender bias in machine translation: a case study with Google Translate*, in *Neural Computing and Applications*, 2020, p. 6363.

[73] C. Nardocci, *Dalla parola che discrimina alla parità nel linguaggio, la dimensione sovranazionale*, in M. Brambilla - M. D'Amico - V. Crestani - C. Nardocci (eds.), *Genere, disabilità, linguaggio. Progetti e prospettive a Milano*, Milano, 2022, p. 53 ss.

[74] A. Akarov, *Did You Just Assume My Vector? Detecting Gender Stereotypes in Word Embeddings*, in W.M.P. Van der Aalst. et al. (eds.), *Recent Trends in Analysis of Images, Social Networks and Texts*, New York, 2021.

[75] Available at https://rare-technologies.com/word2vec-tutorial/#bonus_app, which uses the word2vec model trained by Google on the Google News dataset on about 100 billion words.

[76] Definition provided by Cambridge Dictionary.

Google: UN Women (the United Nations agency for women's rights) exposed the sexist phenomenon of Google's autocomplete function: in particular, when writing on the bar statements such as "Women should…" the suggestions were "stay at home"; "be slave"; "be in the kitchen"; "not speak in the church"; while entering "Women shouldn't…" the results were "have rights"; "vote"; "work"; "box"; and to the question "Women need to…" the answers were "be put in their place", "know their place", "be controlled", "be disciplined". Christopher Hunt, the Art Director of the creative team, stated that: «When we came across these searches, we were shocked by how negative they were and decided we had to do something with them»[77]. Therefore, they had the idea to put the text of the Google searches over the mouths of women portraits, as if to silence their voices[78].

For UN Women, the searches confirm the urgent need to continue making the case for women's rights, empowerment and equality, a cause the organization is pursuing around the world. This was in a second moment echoed by the "Global Voices community", an international network of bloggers and volunteer citizen reporters, who carried out the same research in 12 languages and from different continents, discovering similar conclusions.

One more example came from a recent University of Virginia and University of Washington study[79], which revealed that two image-recognition software programmes, including one supported by Microsoft and Facebook, are keen to associate photos of kitchen with women, even if the one represented in the picture is a man: the result was that the programme labelled a man as a "woman" just because he was at the stoves.

The main damage produced by these kinds of bias is that they are breeding ground for further discriminations, because current software programmes are trained on old ones. Therefore, they learn further associations and they will reproduce the same bias on a larger scale.

Indeed, if major tech companies replicate these altered connections, that will affect the normal function of all home assistants with cameras (like

[77] Reported by J. Yap in the article *These ads show you that sexism is still widespread in the society today available* at: https://vulcanpost.com/2219/these-ads-show-you-that-sexism-are-still-widespread-in-the-society-today/.

[78] The portraits are available at: https://www.unwomen.org/en/news/stories/2013/10/women-should-ads with a brief description of the experiment.

[79] See J. Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*, in *arXiv preprint arXiv:1707.09457*, 2017. The authors' analysis reveals that over 45% and 37% of verbs and objects exhibit bias toward a gender greater than 2:1.

Portal + by Facebook or Amazon Look), the social media targeting and also the automated recruitment process.

With reference to the latter one, nowadays companies use AI assisted procedures to review job applicants' curricula and to preselect potential candidates. As said before, due to the black box-issue, motivations behind the final decision are often hidden[80]. Therefore, issues concerning gender equality in the hiring process might remain unnoticed or very difficult to prove, even though several studies have already shown that women are negatively affected by automated recruitment process[81].

For instance, it has been found that a machine-learning algorithm used by Amazon to select the best candidates for interviews favoured male candidates[82]: indeed, the CVs used for training came from previous applicants who were predominantly male. In this regard, if the data incorporated into the algorithm process lacks population diversity, results will be certainly based on stereotypes, according to the general principle that "if you put garbage in you get garbage out". In this case, given the little number of women hired and working Amazon in the last ten years, the algorithm easily noticed the male supremacy, recognizing in it a factor of success and therefore replicate it in the hiring process[83].

[80] For a complete analysis of the black box issue related AI see to Y. Bathaee, *The artificial intelligence black box and the failure of intent and causation*, in *Harvard Journal of Law & Technology*, 2018, p. 889.

[81] On this topic, see the report of The EU Mutual Learning Program in Gender Equality Artificial Intelligence and Gender Biases in Recruitment and Selection Processes - online seminar - 12-13 November 2020 available at https://ec.europa.eu/info/publications/artificial-intelligence-and-gender-biases-recruitment-and-selection-processes-online-seminar-12-13-november-2020_en. In particular, to have an overview on the Italian situation, it is possible to examine the Comments paper of Italy.

[82] The news was firstly reported by Reuters (J. Dustin, *Amazon scraps secret AI recruiting tool that showed bias against women*, 2018) and further analysed by M. D. Dubber - F. Pasquale - S. Das, *The Oxford Handbook of Ethics of AI*, Oxford, 2020 and L. Devillers et al., *AI & Human Values* in B. Braunschweig et al. (eds.), *Reflections on Artificial Intelligence for Humanity*, New York, 2021.

[83] To have an in-depth examination of some substantial cases concerning discriminatory protection with respect to the use of algorithms within the decision-making processes developed by the public administration or private actors see G. Capuzzo, *"Do Algorithms dream about Electric Sheep?" Percorsi di studio in tema di discriminazione e processi decisori algoritmici tra le due sponde dell'Atlantico*, in *Medialaws*, 2020, p. 102. Since case-study on this field is broad, the Author decided to focus on two categories: the selection of employees and students, and online advertising.

This kind of bias is strictly related to the so-called historical bias, according to which if in the past more males were selected for high position jobs, advertisements of high position jobs will be shown mainly to males[84], generating a tangible disparity of access to information. Therefore, if gender bias affects both automated HR recruitment and HR development – which usually target and hire mostly men – the chances to obtain equality between woman and men in the labour market are increasingly distant[85]. In this sense, discrimination is a real danger: if women in a company have not held managerial positions in the past, they will not even be able to influence the data set that will determine who will do so in the future[86]. In this regard, the workers' advocacy group Upturn published a comprehensive report on the problems with recruitment algorithms, urging that «concrete steps be taken to identify and remove their biases»[87].

In this context, segregation in education and work (namely concentration of women or men in certain jobs) is another serious issue: it has actually increased since 2010 both in EU – where only two out of ten ICT jobs are held by women – and in the US, where notably the top tech companies are dominated by man[88].

The lack of female presence in the workplace has serious consequences, which in the case of AI development are multiplied, because algorithms can spread discriminations on a massive scale at a rapid pace.

[84] A. Köchling - M.C. Wehner, *Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development*, in *Bus Res*, 2020, p. 795.

[85] As shown by the Gender Equality Index developed by the European Institute for Gender Equality (EIGE), with a score of 67.9 out of 100, the EU is at least 60 years away from reaching complete gender equality. EIGE's Gender Equality Index shows that advances in gender equality are still moving at a "snail's pace", with an average improvement of just half a point each year. Gender Equality Index 2020 was acknowledged as a reliable measurement tool for gender equality in the European Union, in an audit carried out by the European Commission's Joint Research Centre.

[86] F. Pasquale, *Le nuove leggi della robotica. Difendere la competenza umana nell'era dell'intelligenza artificiale*, Roma, 2021.

[87] M. Bogen - A. Rieke, *Help Wanted: An Examination of Hiring Algorithms, Equity and Bias, Upturn*, Washington, 2018, available at: https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf.

[88] See J. Dastin, *op. cit.*.

Therefore, bias resounds beyond the workplace into the entire world[89], and part of the negative consequences are extremely evident in the case of voice assistants. They are nowadays everywhere: they help people control their homes, they help workers do quicker their job and they are interlocutors of a large number of people. Indeed, personal digital assistants are at our disposal, always ready to help us solve our problems while tracing our user preferences[90]. Often, they are developed through an anthropomorphizing process in order to create the illusion to be capable of a human dialogue. Although voice assistants seem a new creation, they have been with us since a long time: one of the first chatbots able to perform a natural language process application was introduced already in 1966, with the name (once again feminine) of Eliza. Eliza had the specific aim to act as a psychotherapist, giving responses which made the users feel they were talking to a human who understood their inputs and problems. The resulting bot was designed to undertake real interactions with human based on simulation, so that the virtual therapist thinks about the question by turning back the same question to the patient[91]. Since this first experiment, the world of virtual assistants has been incredibly increased and nowadays they can be divided in general and specialized[92]. The former – like Siri, Cortana, Alexa – is usually integrated into our computers or mobile phones in order to assist us in sending emails, making calls and setting reminders; the latter operates in very specific domains for very specific tasks, like, for instance, SuperFish, which is a language learning chatbot used to teach English at scale[93].

[89] S.M West - M. Whittaker - K. Crawford, *Discriminating Systems: Gender, Race and Power in AI*, AI Now Institute, 2019, available at https://ainowinstitute.org/discriminatingsystems.pdf.

[90] For an interesting analysis on this topic, see P. Costa, *Conversing with personal digital assistants: on gender and artificial intelligence*, in *Journal of Science and Technology of the Arts*, 2018, p. 59.

[91] To have a complete overview on the history of Eliza see C. Bassett, *The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present*, in *AI & Society*, 2019, p. 803.

[92] R. Dale, *The return of the chatbots in Natural Language Engineering*, 2016, p. 811.

[93] As reported from the official website (http://www.superfishai.com/index-en.html), for the last 3 years over 100 thousand students from 450 schools have used SuperFish InteliClass™. It provides a simple solution to rural areas in China where there is a shortage of quality English teachers. According to official report from Chinese government, at least 100,000 English teachers are needed to fill the gap. The company's approach is to employ AI technology to deliver a simple, standardized English learning platform for teachers, students, parents and administrators. To pursue its goal, SuperFish uses a chatbot named Computer-Assisted Language Learning (CALL). CALL software

Mainly focusing on the general ones, it appears that they have been developed through an anthropomorphizing process which results in female gender: they are given human traits, voices and avatars which ended to be female, and this choice was anything but random. The predominance of female AI voice assistants has proven to be strictly connected to the fact they are considered more "humble", "respectful" and "helpful": user research data indicated that people preferred to talk to a female "persona" and that they preferred a woman in a subservient assistant type role[94]. This is a direct consequence of existing gender stereotypes: our society expects women to cover certain type of job, according to their expected personal characteristics: historically women have filled the role of assisting and establishing calls and communications, or the role of nurses or secretaries. This stereotypical image of a women "born to care", to assist or to have a general altruistic behaviour was translated in the AI world, where virtual assistant – which ontologically operates in context of service – are designed to be female, based on the assumption that women possess a natural affinity for service work and emotional labour[95]. At the same time, the strong presence of female virtual assistants has created in the virtual world the same context of violence and abuses present in the real world. Indeed, when it is added a pronoun to the virtual assistant and that pronoun is "she", abusing conversations increased, as well as when it is added a female voice or an avatar with a female face[96].

In this regard, especially in the recent past, often the behaviour of general voice assistant confirms the stereotyped expectations about gender, as they tended to have a submissive "personality". It has been found that some female voice assistants remained impassive in the face of sexist insults

combines speech recognition, artificial speech, and interactive teaching techniques with the convenience of mobile phones and tablets. The effectiveness of CALL software has been demonstrated in numerous academic studies, and its popularity with students, teachers and administrators has also been shown as well.

[94] In this sense, P.L. Frana - M. Klein (eds.), *Encyclopedia of Artificial Intelligence: The Past, Present, and Future of AI*, London, 2021, p. 158; M. D. Dubber - F. Pasquale - S. Das, *op. cit.*, p. 260.

[95] See H. Hester, *Technology Becomes Her*, in *New Vistas*, 2016, p. 46.

[96] A. Piper, *Stereotyping femininity in disembodied virtual assistants*, Graduate Theses and Dissertations 15792, Iowa State University, 2016, p. 58-62. On this topic see also C. Nass et al., *Are Computers Gender-Neutral? Gender Stereotypic Responses to Computers, in Journal of Applied Social Psychology*, 2006, p. 12; J.R. Bookwalter, *Siri Says the Darndest Things: 50 Questions for Apple's Virtual Assistant*, in *Macworld*, available at http://macworld.com/article/2915908/siri-says-the-darnedest-things-50-questions-for-apple-s-virtual-assistant.html, 2015.

or even consider sexist comments as compliments. For instance, in 2017 Siri was used to reply: "I'd blush if I could" to "You are a slut", while Alexa considered it as normal feedback[97].

These shocking results, as well as the predominance of female voice assistants and their subservient personalities, can be attributed to the fact that they are designed by teams that are overwhelmingly male. Indeed, according to the World Economic Forum 2018 Global Gender Gap Report, only 22% of AI professionals globally are female[98]. As mentioned above, if the design stage of AI tools is dominated by man, gender discrimination would be always more likely to be generated: the lack of diversity in design teams causes a lack of democracy in establishing which forms of AI are appropriate and how they can be created and used in ways that respect minority rights. Indeed, the persistent under representation of women in ICT, and in AI in particular, contributed to create a much broader gender segregation in the labour market. In this regard, we can assist to a sort of vicious circle: major startups are funded by Venture capital investment (VC), however only 2% of the startups that receive VC funding are led by female founders[99]. One of the main reasons seem to be connected to the fact that investors are predominantly males, who tend to prefer and invest in males rather than in women while, on the contrary, female VCs are twice as likely to invest in female founders[100]. In a nutshell, if more male-led companies are funded, more male-led companies would have success and

[97] L. Fessler, *Voice Assistant Responses to Sexual or Gender-Based Harassment*, in *Quartz*, 2017.

[98] Global Gender Gap Report 2018, published on 17 December 2018 and available at https://www.weforum.org/reports/the-global-gender-gap-report-2018. Furthermore, only 22% of AI programmers are women. See European Commission, Striving for a Union of Equality, the Gender Equality Strategy 2020-2025, 2020, available at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/gender-equality/gender-equality-strategy_en#gender-equality-strategy-2020-2025.

[99] Crunchbase data show that not only did total funding to female-led startups fall in the last years, but the proportion of dollars to female-only founders also declined, to 2.3 percent, compared to 2.8 percent in 2019 (complete report available at https://about.crunchbase.com/wp-content/uploads/2020/03/Funding-To-Female-Founders_Report.pdf).

[100] In this sense, see the PitchBook-All Raise Report, *All In: Women in the VC Ecosystem, sponsored by Microsoft for Startups and Goldman Sachs' Launch With GS*, 2019, p. 24-25. On this topic, an interesting experiment was carried out by A. W. Brooks et al, *Investors prefer entrepreneurial ventures pitched by attractive men*, in *Proceedings of the National Academy of Sciences of the United States of America*, 2014, p. 4427.

more VCs will choose to fund mainly male-led companies. Therefore, gender stereotypes are reinforced, and woman's representation decrease[101].

An increase in diversity in the AI sector could accelerate solutions to issues related to gender discrimination. To solve or at least reduce this problem, a team of researchers, sound designers and linguists in an initiative called Equal AI developed the first genderless voice assistant, which present itself with the following presentation: «Hi, I'm Q, the world's first genderless Voice Assistant. Think of me like Siri or Alexa but neither male nor female. I'm created for a future where we are no longer defined by gender but rather how we define ourselves. My voice was recorded by people who neither identify as male nor female and then altered to sound gender neutral putting my voice between 145 and 175 Hertz, arranged defined by audio researchers. But for me to become a third option for voice assistance, I need your help share my voice with Apple Amazon Google and Microsoft and together we can ensure that technology recognizes us all. Thanks for listening. Q»[102].

This kind of experiment could surely contribute to reduce discrimination about female gender and bring AI back to its nature: an entity without gender. Indeed, the creator of Q explicitly explained that «Another goal of Q is to give businesses an option to challenge gender stereotypes».

In any case, it is essential to bear in mind that current algorithms exist because there is gender discrimination along the axis of power that allows one gender to prevail over the others. This was (and still is) the model that inspires the architecture of modern artificial intelligences and ICT technologies are its pragmatic realisation. The gender bias is at work in the new forms of algorithmic discrimination in many applications, where the human possibility of establishing the parameters of standards is masked and reduced to a black box, which, once inserted into the interpretative scheme, can make AI racist and sexist[103].

Since bias can be introduced into the algorithm during the data preparation phase, the most delicate moment is the selection of the criteria to be evaluated by the algorithm. In the case of an AI that wants to determine the risk of bank default, for example, one criterion could be the

[101] This analysis was made by P. Fund, in the Annual Meeting of the New Champions (2019) and reproduced in the article *This is why AI has a gender problem*, World Economic Forum, 2019, available at https://www.weforum.org/agenda/2019/06/this-is-why-ai-has-a-gender-problem/.

[102] See the official page of Q at https://www.genderlessvoice.com

[103] In this sense, see M. Vaccari, *Appunti di femminismo digitale #2 Algoritmi*, Independently Published, 2021.

age, income or number of loans already requested by the customer. In the case of an algorithm used to recruit, the criteria might be the candidate's level of education, gender, or years of experience. The choice of which criteria to consider or ignore may significantly influence the predictive ability of the model. The problem is that the impact on prediction accuracy is easy to measure, whereas the impact on polarization is not. Polarization is the "correctness" of the algorithm and is different from precision. For example, if the algorithm predicts that a male candidate would be more likely to be hired, it may be right, but only because the historical data used at the time showed such a trend, being the consequence of an existing gender-bias then replicated by the algorithm.

Many experts point out that the best antidote to the algorithmic vicious circle of prejudice is the adoption of *ad hoc* company policies[104]. Policies should recognize the risk of discrimination and subject algorithms to continuous scrutiny: firstly, by preventing AI from working with potentially discriminatory targets and attributes, and secondly, by preventing it from being trained with biased data. However, the problem is subtle, as underlined by D'Amico: «to make AI and its language more sensitive to the perspective of gender is not enough to rethink datasets by adding or removing data. Overcoming the stereotypes employed by AI technologies is not easy for at least two reasons: AI reflects the biases present in the social fabric; its operation is not easily controllable. Algorithms are the sole property of those who programmed them and, in the most complex cases, programmers are unable to control how such technologies work to correct possible malfunctions»[105].

it may be necessary to scrutinize the output of the algorithm *ex post* to assess whether the results are biased in any way. It is therefore crucial that all links in the chain are aware awareness of this issue: those who develop the algorithms, those who offer them and those who then adopt them in their organizations[106].

[104] Reported by A. Longo - G. Scorza, *Intelligenza artificiale. L'impatto sulle nostre vite, diritti e libertà*, Milano, 2020, p. 126.

[105] M. D'Amico, *Linguaggio discriminatorio e garanzie costituzionali*, in Rivista AIC, 2023, p. 245. The original text is in Italian and has been translated in English by the author.

[106] A. Longo - G. Scorza, *op. cit.*, p. 127.

5. *The role of corporate social responsibility: good and bad practices*

In both the corporate and academic world there is an abundance of definition of social corporate social responsibility (CSR)[107]. To cite just a few, CSR: «is the continuing commitment by business to behave ethically and contribute to economic development while improving the quality of life of the workforce and their families as well as the local community and society at large»[108] or «is the voluntary assumption by companies of responsibilities beyond purely economic and legal responsibilities»[109] or, again, «can be defined as the set of practices and behaviours that firms adopt toward their labour force, towards the environment in which their operations are embedded, towards authority and towards civil society»[110].

Therefore, in a nutshell CSR is an instrument through which companies go beyond mere adherence to legal requirements and instead, incorporate socially responsible practices into their fundamental values. Indeed, this is the concept also embraced by the EU, that sees CSR as: «a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis. It is about enterprises deciding to go beyond minimum legal requirements and obligations stemming from collective agreements in order to address societal needs. Through CSR, enterprises of all sizes, in cooperation with their stakeholders, can help to reconcile economic, social and environmental ambitions. As such, CSR has become an increasingly important concept both globally and within the EU, and is part of the debate about globalisation, competitiveness and sustainability»[111].

[107] To have a critical overview about the numerous efforts to bring about a clear definition of corporate social responsibility see A. Dahlsrud, *How Corporate Social Responsibility Is Defined: An Analysis of 37 Definitions*, in *Corporate Social Responsibility and Environmental Management*, 2008, 1.

[108] R. Holme et. al (World Business Council for Sustainable Development), *Corporate Social Responsibility: Making Good Business Sense*, Geneva, 2000.

[109] M. G. Piacentini et al., *Corporate social responsibility in food retailing*, in *International Journal of Retail and Distribution Management*, 2000, p. 459.

[110] T. Foran, *Corporate Social Responsibility at Nine Multinational Electronics Firms in Thailand: a Preliminary Analysis*, report to the California Global Corporate Accountability Project (Nautilus Institute for Security and Sustainable Development), 2001.

[111] Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee, *Implementing the partnership for growth and jobs: making Europe a pole of excellence on corporate social responsibility*, COM/2006/136 final.

In recent years, there has been a growing trend among companies to actively address the gender equality agenda and including it within their CSR programmes: indeed, it is increasingly recognized that advancements in gender equality in the field CSR can contribute to broader gender and sustainability goals set by the EU[112].

It is also believed that developments in board gender diversity (BGD) have a considerable positive impact on CSR rating, performance, and reporting[113], at the point that the term "gendered social responsibility" (GSR) has been created, referring to the incorporation of gender equality objectives into the social responsibility practices[114].

Therefore, BGD and CSR tend to be interrelated. In this respect, for instance, also the EU has recently taken a step forward improving the gender balance in corporate boards. Already in 2012, the Commission proposed the adoption of the Directive on improving the gender balance among non-executive directors of listed companies and in 2022 a political agreement was reached between the European Parliament and the Council on the point[115].

[112] K. Grosser, *Corporate Social Responsibility and Gender Equality: Women as Stakeholders and the European Union Sustainability Strategy*, in *Business Ethics: A European Review*, 2009, p. 290.

[113] See, for instance, K. Baker et al., *A bibliometric analysis of board diversity: current status, development, and future research directions*, in *Journal of Business Research*, 2020, p. 232, who analyzed how women are more driven toward philanthropic activities, improving firm performance and CSR implementation. See also M. C. Pucheta-Martínez et al., *Corporate governance, female directors and quality of financial information*, in *Business Ethics: A European Review*, p. 363. and S. Escamilla-Solano et al., *Disclosure of gender policies: do they affect business performance?*, in *Heliyon*, 2022, 1.

[114] E. Velasco et al., *Guía de buenas prácticas en responsabilidad social de género*, Madrid, 2014.

[115] «The agreed Directive will ensure that gender balance in corporate boards of listed companies is sought across the EU, while allowing for flexibility for Member States that have adopted equally effective measures. This flexibility will allow for the suspension of the procedural requirements set out in the Directive. The main elements of the Directive are: 1) At least 40% of the underrepresented gender must be represented in non-executive boards of listed companies or 33% among all directors. Member States have to ensure that companies strive to achieve this objective. Those companies that do not achieve those objectives must apply transparent and gender neutral criteria in the appointment of directors and prioritise the underrepresented sex where two candidates of different sexes are equally qualified. 2) Clear and transparent board appointment procedures with objective assessment based on merit, irrespective of gender. The selection procedure of non-executive directors will need to comply to the following binding measures: a) where two candidates of different sexes are equally qualified, preference shall be given to the candidate of the underrepresented sex, in companies where the target for gender balance is not

In this respect, the issue of BGD, CSR and private sector accountability for gender equality appears to be extremely important in the present moment[116]. This relevance arises not just from the European Union's requirement for companies to actively support gender equality efforts but also due to the increasing influence of the private sector in shaping social governance. Indeed, the role of governments as the only source of authority concerning regulation has been transformed[117].

In terms of regulatory sources, there has been a shift away from the monopoly of state legislation, moving towards a system of internal legal pluralism[118]. This includes various forms of regulation such as soft law, contracts, standards, and CSR. These changes have emerged from processes like privatization, which is often described as the "hollowing out" of government (as coined by Rhodes in 1996), or a shift in the balance of governmental roles from "row" to "steer", as articulated by Osborne and Gaebler in 1992.

achieved; b) companies must disclose their qualification criteria should the unsuccessful candidate request it. Companies are further responsible to prove no measures were transgressed, if there is suspicion that an unsuccessful candidate of the underrepresented sex was equally qualified; c) companies must undertake individual commitments to reach gender balance among their executive directors; d) companies that fail to meet the objective of this Directive must report the reasons and the measures they are taking to address this shortcoming; e) member States' penalties for companies that fail to comply with selection and reporting obligations must be effective, proportionate and dissuasive They could include fines and nullity or annulment of the contested director's appointment. Member States shall also publish information on companies' that are reaching targets, which would serve as peer-pressure to complement enforcement ("faming" provision)». See the Press Release of the European Commission available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_3478.

[116] K. Grosser, *op. cit.*, p. 290.

[117] K. Grosser - J. Moon, *Gender mainstreaming and corporate social responsibility: reporting workplace issues*, in *Journal of Business Ethics*, 2005, p. 327.

[118] This transformation has been addressed by C. Scott and described as the rise of the post-regulatory State: «state of mind which seeks to test the assumptions that states are the main loci of control over social and economic life or that they ought to have such a position and role. In the age of governance regulatory control is perceived as diffused through society with less emphasis on the sovereign State. This preliminary investigation of the legal dimension to the post-regulatory State is a long way from asserting the unimportance of law to contemporary regulation. At a descriptive level the analysis offers a wider array of norm-types and control mechanisms relevant to understanding regulatory governance than is common in functionalist analyses of the regulatory State. Normatively the analysis is suggestive of alternative functions for law to asserting command». See C. Scott, *Regulation in the Age of Governance: The Rise of the Post Regulatory State*, in J. Jordana - D. Levi-Faur (eds.), *The Politics of Regulation: Institutions and Regulatory Reforms for the Age of Governance*, Cheltenham, 2004, p. 145 ss.

Within this context, the convergence of the CSR agenda with the increased prominence of gender-related concerns, and the subsequent corporate commitment to advancing the empowerment of women and girls, proves to be of utmost importance[119].

However, the emergence and the dominance of AI introduces a new dimension of concern where the progress made in gender equality policies may face potential setbacks. As mentioned in the previous paragraph, algorithmic discrimination tends to be more nuanced and less conspicuous than human discrimination. Hence, despite the adoption of GSR practices, the use of AI has the potential to give rise to a new type of gender discrimination that should be specifically addressed: AI-driven gender discrimination.

Indeed, there are a wide range of examples which reveal that gender discrimination in AI is nowadays a reality and this reality is particularly evident mainly in the top-tech companies. In addition to the example reported on §4 about recruitment automated processes[120], a recent study has proved that the major facial recognition systems are seriously gender biased. The research conducted by Buolamwini and Gebru examines how well different gender classification systems worked across different people faces and if the results changed based on somebody's gender or their skin type[121].

They created a data set of over 1000 images to have a wide range of skin types and they chose three companies to evaluate: IBM Microsoft and Face ++. With the data set and the companies selected they ran a test: all companies perform better on males than females and all companies also performed better on lighter subjects than on darker subjects. Indeed, the analysis of the results for the four sub-groups showed that all companies performed worst on darker females: as shown in the table, IBM and Microsoft performed best on lighter males and Face ++ performed best on darker males compared to the others. In the majority of cases women are the most discriminated. One of the reasons for this kind of bias is the lack of diversity in training images and datasets failure to separate accuracy

[119] S. Calkin, *Globalizing "Girl Power": Corporate Social Responsibility and Transnational Business Initiatives for Gender Equality*, in *Globalizations*, 2016, p. 158.

[120] It has been found that a machine-learning algorithm used by Amazon to select the best candidates for interviews favored male candidates, since the CVs used for training came from previous applicants who were predominantly male.

[121] J. Buolamwini - T. Gebru, *Gender Shades: Intersectional Accuracy Disparities*, in *Commercial Gender Classification, in Conference on fairness, accountability and transparency*, in *Proceedings of Machine Learning Research*, 2018, p. 77.

results across traits like gender and skin type also makes it harder to identify differences.

However, according to the UNESCO report aforementioned there are a wide range of steps which can be taken in order to reduce gender discrimination in the social responsibility context, starting by setting up adequate measures for their prevention, mitigation and remediation[122]. Indeed, companies have the responsibility to ensure the respect of human right to non-discrimination, since business responsibility exists independently of States abilities to create a national normative protection[123].

First of all, companies should enhance company governance models and mechanisms for ethical compliance, including reflections on gender equality and the involvement of women. They should create highest-level policies that show strong management support for advancing gender equality through corporate products and services. In this regard, an efficient step could be the creation of incentives for non-biased products, such as promotions or other benefits[124].

In 2017, some of the major American companies – including Google, Microsoft, Facebook and Amazon – started the Partnership on AI to Benefit People and the Society (PAI): this is one of the biggest groups composed by AI engineers working with non-experts to elaborate best practices and guidelines in different areas of AI[125]. PAI last annual report (2020) – in the Equity&Inclusion Section – focused on the lack of diversity in the field of AI, announcing that in July 2020 PAI formalized its commitment to investigating the AI's diversity gap with the hire of a

[122] UNESCO, *op. cit.*

[123] In this sense, the United Nations Guiding Principles on Business and Human Rights (UNGPs) provided some general principle, among which principle 11: «Business enterprises should respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved». The commentary is available at https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

[124] UNESCO, *op. cit.*, pp. 21-23.

[125] On the official website of the foundation, which is consultable at https://www.partnershiponai.org/about/, there is the explicit goals list: «first, to develop and share best-practice methods and approaches in the research, development, testing, and fielding of AI technologies; second, to advance public understanding of AI across varied constituencies, including on core technologies, potential benefits, and costs; third, to provide an open and inclusive platform for discussion and engagement on the future of AI, and to ensure that key stakeholders have the knowledge, resources, and overall capacity to participate fully in these important conversations; and fourth, to identify and foster aspirational efforts in AI for socially benevolent applications».

Diversity and Inclusion Fellow[126]. In this regard, PAI research launched a study with the aim to investigates high attrition rates among women and minoritized individuals in tech, since all organizations that use AI are increasingly aware of the need to actively challenge bias in the products they produce[127].

Therefore, it seems that something is changing: looking at individual realities, there are examples of company governance models which include reflections on gender equality and the involvement of women. For instance, some companies succeed in drafting AI principles and creating guidelines capable of reducing gender inequality, by choosing to use datasets that have been developed with a gender equality lens (despite incurring in higher costs). Others have created *ethical advisory council* with experts and civil society groups, which integrates gender equality concerns and ensure an interdisciplinary approach.

In this regard, for example, Microsoft launched an internal high-level group on AI and Ethics in Engineering and Research (AETHER) which examines how its software should or should not be used[128]. The AETHER Committee was created in 2016: it serves an advisory role to the company's senior leadership on rising questions, challenges, and opportunities with the development of AI technologies. The Committee works on specific topics creating different working groups, which include teams focused on bias and fairness, intelligibility and explanation, facial recognition systems and reliability and safety. The principles and the best practices enucleated during the research have been in 2018 collected into the official document «The Future Computed. Artificial Intelligence and its role in society»[129].

[126] Partnership on AI Annual Report, 2020, p. 7-8, available at https://www.partnershiponai.org/annual-report-2020/. The Partnership on AI (PAI) is an independent, nonprofit organization. It was originally established by a coalition of representatives from technology companies, civil society organizations, and academic institutions, and supported originally by multi-year grants from Apple, Amazon, Facebook, Google/DeepMind, IBM and Microsoft.

[127] *Ibid.*

[128] The AETHER Committee has different working groups on specific topics. Working groups include teams focused on sensitive uses of AI, bias and fairness of AI systems, intelligibility and explanation of AI reasoning and recommendations, reliability and safety, human-AI interaction, and engineering best practices. As explained in the official site, AETHER also provides guidance to teams across the company to ensure that AI products and services align with Microsoft's AI principles. The committee includes top talent in research, engineering, ethics, law, and policy from across Microsoft who come together to formulate recommendations on policies, processes, and best practices.

[129] Available at https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/The-Future-Computed_2.8.18.pdf.

However, in the preamble of this document, while describing how our morning will be in twenty years by now, it is still present the pronoun "she": «At Microsoft, we imagine a world where your personal digital assistant Cortana talks with your calendar while you sleep. She works with your other smart devices at home to rouse you at the end of a sleep cycle when it's easiest to wake and ensures that you have plenty of time to shower, dress, commute and prepare for your first meeting (…). A digital assistant like Cortana will then automatically prepare a summary of the meeting with tasks assigned to the participants and reminders placed on their schedules based on the conversation that took place and the decisions the participants made».

It seems then than even if the path has been opened, it will still take a long time to reach the destination. Nevertheless, at least nowadays there are adequate instruments to deal with this journey: companies developing and using machine learning system have started to bring principles of AI non-discrimination to life, by ensuring practices of proactive due diligence. The World Economic Forum's White Paper on "How to prevent discriminatory outcomes in Machine Learning" indicated three main steps which have to be followed by every company involved in the use of AI:

- Identifying human rights in business context;
- Taking effective measures to prevent or mitigate risks.
- Being transparent about efforts to identify, prevent and mitigate human rights risks[130].

What is still missing is an autonomous category of tools specifically designed to counter gender discrimination and this gap is not only present in social corporate responsibility context, but also at the legislative level, as already outlined above.

However, action to recognize and mitigate bias in AI is being taken by different stakeholders: companies, academia, non-governmental organizations (NGOs), and – as reported by the Playbook "Mitigating Bias in AI" created at Berkeley – also the Roman Catholic Church[131]. The

[130] The White Paper was published on the 12th March 2018 and it is available at https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning.

[131] G. Smith - I. Rustagi, *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook*, Center for Equity, Gender and Leadership, Berkeley Haas, 2020, reported that «The Roman Catholic Church joined with IBM and Microsoft to work on ethics of artificial intelligence in the "Rome Call for Ethics". The call, which outlines three principles, was supported by Pope Francis who made detailed remarks about the impact of AI on humanity». (https://romecall.org/).

Playbook identified precise steps to follow in order to mitigate bias: first, to enable diverse and multi-disciplinary teams working on algorithms and AI system and to promote a culture of ethics and responsibility related to AI. These to step are categorized under the section "Team", since they directly concern the staff and serve to enable a culture that encourages the employees themselves to prioritize equity considerations at every step of the algorithm development process[132].

Second, to practice responsible dataset development and to establish policies and practices that enable responsible algorithm development. In this regard, a company should ask itself: «Do teams developing machine learning datasets assess the quality and quantity of data generated and gathered to ensure population is sufficiently and accurately represented?»; «Are there robust feedback mechanisms built into AI systems so users can easily report performance issues they encounter, and (if no way to opt out), have an appeal process to request human review?»[133].

Lastly, to build a leadership able to establish corporate governance for responsible AI and end-to-end internal policies to mitigate bias. In this sense, the company should have an AI ethics lead and AI ethics board, in addition to an AI ethics code, and should engage corporate social responsibility to advance responsible AI. To this aim, it would be crucial to leverage corporate social responsibility teams, creating different incentive structure to advance responsible AI internally than other parts of the business with less priority on efficiency.

The Playbook indicated also a final step: «to use your voice and influence to advance industry change and regulations for responsible AI». This means engaging in partnerships with various stakeholders to advocate for policies for responsible AI and approaches in industry. In this regard, it is important to fund research to advance knowledge and to create the possibility to operate a diversification between research teams.

In putting in practice these indications, it is crucial to keep in mind that the final goal is not – and cannot be – the full "de-biasing" of AI[134]. Only a mitigation is possible and is not sufficient to realize it with technical solutions alone: to understand and address bias a prerequisite is understand

[132] G. Smith - I. Rustagi, *op. cit.*, p. 10.

[133] G. Smith - I. Rustagi, *op. cit.*, p. 11.

[134] As stated by C. D'Ignazio, Assistant Professor at Massachusetts Institute of Technology (MIT): «Data is never this raw, truthful input and never neutral. It is information that has been collected in certain ways by certain actors and institutions for certain reasons» in Corbyn, Z., Interview: Catherine D'Ignazio, 21 March 2020, The Guardian.

the social, economic, and political context in which the data was produced. Ignoring the background and claiming that data is objective creates a fence to mitigate it: «Note that in order to train an algorithm to understand the context of subjugated standpoints, significant human infrastructure and ethical navigation is required»[135]. Indeed, the context comes into play not just in the stages of data acquisition, but also in the selection and communication of the numbers.

In this sense, to achieve – or at least put some solid ground for a non-gender biased AI – is time to start thinking about a data ethics informed by the idea of intersectional feminism[136]. As argued by Crenshaw already in 1989, focusing solely on either racial discrimination or gender discrimination overlooks the discrimination experienced by black women, contributing to their marginalization in legal and social contexts[137]. To use her words «the intersectional experience is greater than the sum of racism and sexism and any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordined»[138].

Indeed, also within the different categories used by the algorithms, the intersection of multiple identities can exacerbate disadvantages: for instance, of a black woman in the job market can experience discrimination in relation to other women because of their race and at the same time may be discriminated in relation to other men because of their gender. Moreover, even the standard structure of data (that impose exclusionary definitions of identity), could provoke serious marginalization of other categories not included in the binary labels (for instance, queer, transgender, and non-binary people)[139].

The treatment and the study of intersectionality should therefore be incorporated in the analysis and to do that D'Ignazio and Klein suggest having more theory, context, and scientific method in the investigation of

[135] C. D'Ignazio - L. Klein, *Data Feminism*, Cambridge, 2019, p. 103.

[136] C. D'Ignazio - L. Klein, *op. cit.*, Introduction.

[137] K. Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, in *University of Chicago Legal Forum*, 1989, p. 139 ss.

[138] K. Crenshaw, *op. cit.*, 1989, p. 140.

[139] On this topic see O. Keyes, *The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition*, in *Proceedings of the ACM on Human-Computer Interaction*, 2018, p. 1-22.

the human reality in order to interrogate data as they are: «cultural indicators of the changing face of patriarchy and racism»[140].

6. *Concluding remarks*

Nowadays it is almost impossible for companies and individuals to avoid AI systems: therefore, it is urgent to understand how to create a fair AI and to this end it is necessary first to detect the already existing discriminations and exclusions. Although we cannot always have total control on inductive bias that generated by AI, it is always possible to work on the dataset. The creation of a common database with data from all the largest companies using AI could help to detect the existing bias in order to correct them.

As pointed out in the Villani Plan drafted in France, this kind of database will enable «year-to-year progress to be measured and provide a course of action for public policies»[141].

It becomes always more and more evident that business must actively participate to the fight against gender bias in AI, providing their data related to gender balance rate in appointments, promotions and recruitment; gender balance rate in executive committees and board of directors; pay gap between different jobs, at different grades; gender balance rate in teams; and gender balance rate in terms of grade and job type.

Indeed, as suggested also by Buolamwini, it is time to start thinking about how to create more inclusive code. The goal should be creating full spectrum teams with diverse individuals who can check and balance each other. Therefore, the preliminary questions that everybody should ask themselves in the creation and use of AI are: who codes? How they code? And why they codes? It goes without saying that in the last answer gender equality and social change should be a primary reason[142].

In order to achieve gender equality in an AI-driven world, in addition to legislative actions, a suite of independent or complementary policy measures can be implemented to address the concerns highlighted above.

[140] D'Ignazio - L. Klein, *Data Feminism*, *op. cit.*, p. 102.

[141] See the Villani Plan, available at https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

[142] See J. Buolamwini, *How I am Fighting Bias in Algorithms*, presented during a TEDx talks available at https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms.

One effective approach involves targeted training for developers during the algorithm design and coding phases, specifically focusing on gender equality[143]. While this may not eliminate all biases, it substantially diminishes the potential for biases, stereotypes, and discriminatory behaviour in algorithms[144]. In addition to promoting training, mandatory measures at the design and coding stages could be prescribed.

The responsibility for achieving these objectives could be entrusted also to companies, subject to defined and specific legal standards. Concrete outcomes need not be rigidly prescribed by law since creating an entirely bias-free algorithm is often an unattainable goal. The primary objective, therefore, is not the elimination of bias but rather the reduction of the risk associated with gender-based discrimination[145].

If we don't manage to create an ethical an inclusive AI with this specific aim, the risk to lose and violate civil rights and gender equality gains under the illusion of a neutral AI will become a reality.

***

[143] F. Lütz, *op. cit.*, p. 47.

[144] In this sense, it is worth noticing that in 2023 the EU has launched a campaign that aims to raise awareness about the role gender stereotypes play in society available at: https://end-gender-stereotypes.campaign.europa.eu/work-life-balance_en.

[145] F. Lütz, *op. cit.*, p. 48.

ABSTRACT: AI-driven gender discrimination is a new multifaced problem that has the potential to perpetuate and replicate outdated conception of gender roles that society is actively working to eliminate from the "real" world.

This paper identifies gender inequalities resulting from the use of AI and investigates which role legislation, with particularly regard to the forthcoming EU AI Act, and corporate social responsibility (CSR) can play in addressing this issue.

The results of this research underscore the necessity of a multifaceted approach to combat AI gender discrimination. Legislative measures are essential but may fall short in addressing the hidden biases present in AI systems. Policy measures and CSR initiatives are equally vital to complement regulatory efforts, ensuring that AI technologies are developed, implemented, and used without perpetuating gender-based discrimination. In a world increasingly reliant on AI, this holistic approach turns out to be crucial.

***

KEYWORDS: AI-driven gender discrimination – algorithmic bias – EU AI Act –gender equality – corporate social responsibility

***

Federica Fedorczyk – PhD Researcher presso la Scuola Superiore Sant'Anna, Istituto Dirpolis, Pisa (federica.fedorczyk@santannapisa.it)