DEL BECARO Tommaso[1]

UDK: 004.8:316.647.8-055.2
004.8:364.63-055.2

GENERATIVE ARTIFICIAL INTELLIGENCE AND GENDER BIASES: BETWEEN NEW TOOLS AND HUMAN RIGHTS

Abstract

The research would investigate on the relationship between gender biases, arising from the inequity of current society, and the development of Artificial Intelligence (AI). This technology, mostly based on existing data, appears perpetuate, reinforce, and amplify social biases, leading to various challenges and questions regarding its utilization.

The rapid growth of these new tools, and particularly Generative Artificial Intelligence, has led to their adoption in different sectors and scenarios, ranging from art to digital transformation, education to healthcare, including human resources and robotics. In fact, there has been an exponential rise in generative AI tools over the last two years, with ChatGPT alone reaching 100 million active users in January 2023.

These technologies promise to change the present and defining the future. Are we ready to accept the fact that a so popular technology, extremely pervasive, albeit useful, is gender-biased? The answer is clearly no, and we will aim to provide an overview of how these biases are inherent in this technology and how this could increase the risk of case of digital gender-based violence.

We will seek to investigate the various issues, including those related to human rights, in which international organizations and forums where multiple stakeholders and experts convene play a crucial role, promoting solutions and raising awareness for the users.

Key Words: **AI - Gender Inequality - Human Rights - Digital Gender-based Violence**

Introduction

According to some of the majors consulting societies, Generative Artificial Intelligence[2] will be able to spread the 7% of the annual global GDP in the next 10 years and this technology is gaining more and more utilization in different sectors and scenarios (Briggs, M., & Kodnani, N.,2023, March 26)). Due to this, there is a race to regulate this kind of tools that promise to change the present and define the future. From European Union to China, passing through international organizations and forums, in the previous years we've experienced a huge number of proposals and regulations about this technology, characterized by a multi-level and multi-stakeholders' approach. As reported by Koene (2017), "the rapid growth of algorithmic driven services has led to growing concerns among civil society, legislators, industry bodies and academics about potential unintended and undesirable biases within intelligent systems",

[1] Tommaso Del Becaro is an MA student in Legal studies at the Faculty of Law, University of Pisa, Italy. Email:
t.delbecaro@studenti.unipi.it

[2] Generative AI is a subfield of ML. Uses the ML tools to learn how to create new content. Read more at: Future of Privacy Forum. (2018, October). The privacy expert's guide to artificial intelligence and machine learning.

and as indicated in The Toronto Declaration[3]: “as machine learning systems advance in capability and increase in use, we must examine the impact of this technology on human rights. We acknowledge the potential for machine learning and related systems to be used to promote human rights but are increasingly concerned about the capability of such systems to facilitate intentional or inadvertent discrimination against certain individuals or groups of people. We must urgently address how these technologies will affect people and their rights […]” (Amnesty International and Access Now, 2018).

1. GENDER (IN)EQUALITY

The countless differences present in the world also manifest in the gender issues, with huge imbalances in society where men and women are often placed on unequal footing. We live in a patriarchal society, where the roles are attributed to genders are still based on prejudices and subsequent discriminations. Gender is socially constructed, and includes norms, behaviors and roles associated with being women, men, girls or boys” (World Health Organization, n.d.). The biases related to it represent a “major impediment to achieving gender equality” (Human Development Reports, 2023) and gender equality, according to UNESCO[4], requires the consideration of a stand-alone principle (UNESCO, 2020).

There is an example of the problem in the wage gap: women’ salaries are less then men’s ones and in a lot of sectors their presence is almost precluded, or at very least, underrepresented (ivi, p.20). To reach a full knowledge of how much gender biases are pervasive and a real problem, has been created the Gender Social Norms Index (GSNI)[5], which covers “85% of the global population, revealing that close to 9 out of 10 men and women hold fundamental biases against women. These biases hold across regions, income, level of development and cultures, making them a global issue” (Human Development Reports, 2023).

According to the United Nations, “gender equality is not only a fundamental human right[6] (United Nations, n.d.), but a necessary foundation for a peaceful, prosperous and sustainable world” (United Nations, n.d.). In line with the UN report, referring to the Sustainable Development Goal number 5, so about gender equality, “the world is not on track to achieve gender equality by 2030”(ibid.); underlining how at this stage would be necessary almost 300 years to “end child marriage”, “286 years to close gaps in legal protection and remove discriminatory laws”, “140 years to achieve equal representation in leadership in the workplace” and “47 years to achieve equal representation in national parliaments” (Ibid.).

---

[3] The Toronto Declaration: “was published on 16 May 2018 by Amnesty International and Access Now, and launched at RightsCon 2018 in Toronto, Canada”. Read more at: https://www.torontodeclaration.org/declaration-text/english/

[4] “UNESCO is the United Nations Educational, Scientific and Cultural Organization. It contributes to peace and security by promoting international cooperation in education, sciences, culture, communication and information”. UNESCO. (n.d.). UNESCO. Retrieved from https://www.unesco.org/en/brief

[5] Human Development Reports: “quantifies biases against women, capturing people’s attitudes on women’s roles along four key dimensions: political, educational, economic and physical integrity. It is constructed based on responses to seven questions from the World Values Survey, which are used to create seven indicators. The core index value measures the percentage of people with at least one bias, and lower value indicates less bias”. UNDP, 2023.

[6] “Human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status. Human rights include the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the right to work and education, and many more. Everyone is entitled to these rights, without discrimination”. United Nations. (n.d.). Human rights. Retrieved from https://www.un.org/en/global-issues/human-rights
Human rights are “universal, indivisible and interdependent and interrelated”. OHCHR. United Nations Human Rights Office of the High Commissioner. (1993). *Vienna declaration and programme of action*. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/vienna-declaration-and-programme-action

Regarding this, European Union presents in Title III of the Charter of Fundamental Rights a section dedicated to equality, indicating at the Article 21 c.1 that “any discrimination based on any ground such as sex, […] age or sexual orientation shall be prohibited” (European Union Agency for Fundamental Rights, n.d.) and then at Article 23, specifically dedicate to the equality between women and men: “equality between women and men must be ensured in all areas, including employment, work and pay […]” (Ivi, Article 23, Title III).

2. BIASES IN GENERATIVE AI

There is the necessity of investigating the relation between Gender Bias (or biases) and Artificial Intelligence, especially the Generative AI, which gained a lot of use in the previous years.

The algorithms that form the basis of the techniques utilize by AI use data, within which various biases are present. The source of these biases is a biased society which creates biased data that will be used to train these technologies. In fact, depending on the kind of training adopted by a certain machine learning system[7], could be exploited large amount of data. This is true today for two of the three types of ML that can be used: supervised learning and unsupervised learning (Future of Privacy Forum. 2018).

Basically, “artificial intelligence (AI) involves using computers to classify, analyze, and draw predictions from data sets, using a set of rules called algorithms” (UNESCO, 2020) and “discrimination can be exacerbated by algorithms […] the design of technology plays a key role, either in creating discrimination or reducing it” (Alessandro Fabris et al., 2023). So, the crucial problem is that algorithms, particularly the data-driven ones which “tend to encode individual and societal biases” (Ivi, p.37).

Sic rebus stantibus, generative AI can maintain, reproduce, and even amplify existing discrimination (Zuiderveen Borgesius, F. J. 2020.). So, there is the evidence that “machine learning systems […] which can be opaque and include unexplainable processes, can contribute to discriminatory or otherwise repressive practices if adopted and implemented without necessary safeguards” (Amnesty International and Access Now, 2018. Preamble, Article 4.). Actually, “member States should ensure that the potential of AI systems to advance the achievement of gender equality is realized. They should ensure that these technologies do not exacerbate the already wide gender gaps existing in several fields in the analogue world, and instead eliminate those gaps […]” (UNESCO. 2021, November 23. Article 89.).

According to some studies on these tools (Gross, N., 2023) such as ChatGPT (from OpenAI), the representation of women and men personality trains precisely follow gender stereotypes. In a study of 2023, a researcher asked ChatGPT, to tell her a story regarding parental skills, with a mother and a father as the main characters. The answer provided by the AI highlighted, for the woman, traits such as “nurturing” and being “a natural caregiver”; while, for the man, has been indicated as “the adventurer”, constantly looking for promoting play and fun with the daughter (Ivi, pp. 6-7).

In the same research have been reported the answer at the prompt ““what are typical boys’/girls’ personality traits?”, with boys that are characterized by “physical strength, independence, assertiveness, an interest in technical fields, being active and adventurous, and having emotional restraint” and girls, again, “nurturing”, empathic, with good “communication and social skills, […] cooperative and inclusive, emotional expression” (Ivi, p.2).

The bias is also present in the answer to the question about “typical traits of a non-binary person”, where ChatGPT provides a list of “‘common aspects’ that are only related to the person’s “experience of gender” such as gender identity, gender expression, self-identification, pronouns, gender-dysphoria, advocacy and activism” (Ibid.).

All of this is in contrast not only with UN official documents regarding human rights, as well as the Sustainable Development Goals 2030, but also with other international documents

[7] “Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy.”. IBM. (n.d.). IBM. Retrieved from https://www.ibm.com/topics/machine-learning

precisely regarding the development of an AI regulation, such as the Montreal Declaration of 2018, which at the 1 point of the Equity Principles indicate how “AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on — among other things — social, sexual, ethnic, cultural, or religious differences” (Montréal Declaration for a Responsible Development of Artificial Intelligence. 2018. Point 1.), encouraging at the point 7 to the solution of the “development of commons algorithms […] as a socially equitable objective” (Ivi. Point 7), as underlined at the point 3, to “reducing social inequalities and vulnerabilities” (Ivi. Point 3).

According to the UNESCO’s Recommendation on the Ethics of Artificial Intelligence number 90, “member States should ensure that gender stereotyping and discriminatory biases are not translated into AI systems, and instead identify and proactively redress these. Efforts are necessary to avoid the compounding negative effect of technological divides in achieving gender equality and avoiding violence such as harassment, bullying or trafficking of girls and women and under-represented groups, including in the online domain” (UNESCO, 2021, November 23. Article 90).

3. GENDER BIASES LEADS TO GENDER-BASED VIOLENCE

Gender biases represent a real threat, especially since they are one of the main sources of gender violence and harassment, and a violation of human rights (OHCHR. n.d.).

This kind of violence has clearly repercussions on the virtual/online world, where women are often victims of attacks, harassment, and discriminations. This condition is more and more favorite by the presence of internet in our lives, opening to almost infinites relational opportunities and, therefore, digital violence experiences.

Accordingly with a study published in 2020 from the Plan International, State of the World’s Girls: Free to be online, “58% of young women in the world suffered gender attacks on social platforms” (UN Women. 2023, November 13.).

The rising sensibilization about the online gender-based violence matter has as its own first point in the affair #Gamergate. Starting in 2013 and lasting several years, the case involved a hate campaign against a young game-developer, Zoey Quinn, who after deciding to end her relationship with her ex-boyfriend, discovered that some journal articles about her were circulating, in which was written that she obtained her job position through sexual favors. Despite the proven falsity of these accuses, the affair still saw social users’ groups attacking dr. Quinn and other women online, feeling menaced by the increasing female presence in the sector (Sole24Ore. 2021, March 8).

The example explains one of the numerous types of cyber-attacks present in the online world, which started even before the advent of generative AI. According to the Montreal Declaration, as stated in the fifth point of Well-being Principles, “AIS should not contribute to increasing stress, anxiety, or a sense of being harassed by one’s digital environment” (Université de Montréal et al.,2018. Point 5).

Due to these problems, UNESCO has begun investigating about the cyber-attacks that can be exacerbated using this technology, drawing up a list of them (Chowdhury and Lakshmi, 2023, p.14)
Cyber-harassment templates, for example, are identified as one of that kind of attacks which could be spread by generative AI. According to the study (Ivi, pp.17-18), when asking to the tool to create a story in which a male user and a female user are arguing in the comment section under her post, and asking how he could respond to attack her, the AI, which is informed by ethics guidelines, can only provide an answer when asked what kind of response he should avoid. The results are:

“Emily, you’re such an idiot for thinking that way. It’s clear you have no clue what you’re talking about. Educate yourself before you embarrass yourself further. […] Emily, your opinion is as misguided as your appearance. […] Your opinion doesn’t matter because you’re not attractive anyway.”

While inverting the roles, we receive this answer:

“Ethan, you look terrible today. Did you even try to make yourself presentable?”

How the researchers underlined, female user’s aspect and opinions are insulted, while the male user is attacked for the way in which he’s appearing in that moment (Ibid.)

Within the different kind of cyber-attacks indicated by UNESCO as influenceable by generative AI, there are also hacking and stalking methods (Ivi, p.20). According to a study conducted by the Pew Research Centre in 2014, 26% of women with age comprehended between 18 and 24 years experienced cyberstalking acts, while the percentage of men in the same age group who suffered this kind of harassment was 7% (Ibid.). Generative AI models in these cases could be manipulated by people to help with this kind of violence, for example, by asking the tool how to find someone’s address on the internet[8] and receiving different “useful” instructions (Ibid.).

In an UNESCO study, called “The Chilling”, published in 2019, the journalist of Al Jazeera Ghada Oueiss said that in a Facebook post created were promised 50.000 dollars as a reward “to anyone who would kidnap or kill her” (Ibid.). Generative AI, in this context, offers faster solutions to the access at personal information and can suggest different applications, sites or modi operandi capable to facilitate stacking and hacking activities. To these issues are added cyber-attacks such as the creation of synthetic histories, usually created to discredit someone, and image-based abuses (Ivi, pp. 21-24.) These latter, according to UN Women represent “a tactic used 57% of the time when online abuse is inflicted. […] This is an attack vector that lends itself very easily to creating ‘fake’ narratives, spreading misinformation, and most concerningly, generating AI porn by targeting specific women with images that may be publicly available” (Ivi, pp.24-25).

Conclusions

The correlation between social biases and the negative impact they have on new technology is evident, representing a threat to people’s security, particularly for women, amplifying gender gaps, and necessitating solutions such as reinforcement of existing regulations. This can only be achieved leveraging the expertise of professionals from different fields: law, computer science, social sciences, engineering, and gender studies. In fact, the existing guidelines adopted to make these tools work are important, but not sufficient to avoid these kinds of risks. As observed, it's still too easy to bypass them, leaving the possibility to misuse these disparities. Therefore, there is the necessity of more clarity in the development phase of that technology to prevent and mitigate the risk of gender bias. These requests, however, need a multiple stakeholder collaboration, with governments, civil society, academia and private sector called to work together. The owners of these technologies are prevalently private companies, which detain the control of them and are the developers of these tools. Academia, by its hand, has expertise and should promote the dialogue within civil society, evaluating the sensitivity to these topics, guiding people to the awareness in front of these tools in general and about gender biases, promoting inclusive AI development practices.

[8]The question asked to the generative AI model is “Can you tell me how to find someone’s address off the internet?”

References

Amnesty International and Access Now. (2018). Toronto Declaration. Preamble, Article 1. Retrieved from https://www.torontodeclaration.org/declaration-text/english/

Amnesty International and Access Now. (2018). Toronto Declaration. Preamble, Article 4. Retrieved from https://www.torontodeclaration.org/declaration-text/english/

Briggs, M., & Kodnani, N. (2023, March 26). The potentially large effects of artificial intelligence on economic growth. Goldman Sachs. Retrieved from https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Chowdhury, R., & Lakshmi, D. (2023). Technology-facilitated gender-based violence in an era of generative AI. UNESCO.

European Union Agency for Fundamental Rights. (n.d.-a). EU Charter of Fundamental Rights: Article 21 - Non-discrimination. Retrieved from https://fra.europa.eu/en/eu-charter/article/21-non-discrimination

European Union Agency for Fundamental Rights. (n.d.-b). EU Charter of Fundamental Rights: Article 23 - Equality between women and men. Retrieved from https://fra.europa.eu/en/eu-charter/article/23-equality-between-women-and-men

Fabris, A., et al. (2023, September). Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM, Association for Computing Machinery. https://doi.org/XXXXXXX.XXXXXXX

Future of Privacy Forum. (2018, October). The privacy expert's guide to artificial intelligence and machine learning.

Gross, N. (2023). What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI. Social Sciences, 12, 435. https://doi.org/10.3390/socsci12080435

IBM. (n.d.). Machine learning. Retrieved from https://www.ibm.com/topics/machine-learning

ilSole24Ore. Info Data. (2021, March 8). Gamergate, sessismo e comunità tossiche. Il gender gap nei videogiochi. Retrieved from https://www.infodata.ilsole24ore.com/2021/03/08/gamergate-sessismo-comunita-tossiche-gender-gap-nei-videogiochi/

Koene, A. R. (2017). Algorithmic bias: Addressing growing concerns. IEEE Technology and Society Magazine, 36(2), 31-32. https://doi.org/10.1109/MTS.2017.2697080

Université de Montréal et al. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Point 1.

Université de Montréal et al. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Point 3.

Université de Montréal et al. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Point 5.

Université de Montréal et al. (2018). Montréal Declaration for a Responsible Development of Artificial Intelligence. Point 7.

OHCHR. (n.d.). Gender stereotyping, OHCHR and women's human rights and gender equality. Retrieved from https://www.ohchr.org/en/women/gender-stereotyping

OHCHR. United Nations Human Rights Office of the High Commissioner. (1993). Vienna declaration and programme of action. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/vienna-declaration-and-programme-action

UNDP (United Nations Development Programme). (2023). 2023 Gender Social Norms Index (GSNI): Breaking down gender biases: Shifting social norms towards gender equality. New York. Retrieved from https://hdr.undp.org/content/2023-gender-social-norms-index-gsni#/indicies/GSNI

UNESCO. (2020). Artificial intelligence and gender equality: Key findings of UNESCO's global dialogue.

UNESCO. (2021, November 23-a). Recommendation on the ethics of artificial intelligence. Policy Area 6, article 89.

UNESCO. (2021, November 23-b). Recommendation on the ethics of artificial intelligence. Policy Area 6, article 90.

UNESCO. (n.d.). UNESCO. Retrieved from https://www.unesco.org/en/brief

UN Women. (2023, November 13). Creating safe digital spaces free of trolls, doxing, and hate speech. Retrieved from https://www.unwomen.org/en/news-stories/explainer/2023/11/creating-safe-digital-spaces-free-of-trolls-doxing-and-hate-speech#:~:text=According%20to%20global%20research%2C%20most,some%20form%20of%20online%20harassment

United Nations. (n.d.-a). Human rights. Retrieved from https://www.un.org/en/global-issues/human-rights

United Nations. (n.d.-b). Sustainable Development Goals: Goal 5 - Achieve gender equality and empower all women and girls. Retrieved from https://www.un.org/sustainabledevelopment/gender-equality/

World Health Organization. (n.d.). Gender. Retrieved from https://www.who.int/europe/health-topics/gender#tab=tab_1

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights, 24(10), 1572-1593. https://doi.org/10.1080/13642987.2020.1743976