Towards a Gendered Innovation in AI

Silvana Badaloni[a] and Francesca A. Lisi[b]

[a] *University of Padua, Via 8 Febbraio 2, Padua, 35122, Italy*
[b] *University of Bari "Aldo Moro", Via E. Orabona 4, Bari, 70125, Italy*

Abstract

In this paper we address the problem of including the gender dimension in the content of Computer Science, notably in Artificial Intelligence (AI). We analyze first the fairness of Machine Learning (ML) algorithms from a gender point of view. Due to their nature of being bottom-up data-driven algorithms, the most common biases diffused in society about gender and ethnicity can be captured, subsumed and reinforced by them, as many ML applications show. Then, to understand how to develop a new gendered (Computer) Science and promote a gendered innovation in AI, we show a formal reflection on the scientific method utilized to produce innovation and a critical analysis of the logical rules underlying it.

Keywords 1

Gender issues, Bias, Fairness, Trustworthy AI.

1. Introduction

The gender, diversity and inclusion dimension of science and technology has become a highly visible and debated theme worldwide, impacting society at every level. In some fields of knowledge, however, these issues are still not so impactful.

If the term 'AI for good' is increasingly used in the scientific and technological context, there is much less discussion about 'AI for social good' aiming at identifying the relationship between AI and our societal goals, in particular, the goal of *gender equality* [1].

In the field of AI many case studies show that the Machine Learning (ML) algorithms present an "unfairness" from the gender point of view. The hypothesis is that these algorithms are not gender neutral due to their nature of being bottom-up data-driven. They can capture and subsume the most common biases diffused in society and even reinforce them, where for gender bias we adopt the definition given by EIGE [2], i.e. prejudiced actions or thought based on gender-based perception that women are not equal to men in rights and dignity.

In the perspective of developing a *trustworthy AI* able to learn fair AI models even in spite of biased data, as we will illustrate later, we intend to address the problem of framing the landscape of gender equality and AI, trying to understand how AI can overcome gender bias and showing how an interdisciplinary analysis can help in a re-calibration of the biased instruments. This problem is even more important now since AI is often confused with tools, algorithms and technologies developed in its framework [3].

A recent UNESCO report on this subject [4] recognizes the absolute centrality of this topic and provides recommendations on how to address gender equality considerations in AI principles. The purpose of the UNESCO's Dialogue on Gender Equality and AI identifies issues, challenges, and good practices to help:

- Overcome the built-in gender biases found in AI devices, data sets and algorithms;
- Improve the global representation of women in technical roles and in boardrooms in the technology sector;

AIxIA 2020 Discussion Papers Workshop
EMAIL: Silvana.Badaloni@unipd.it (A. 1); Francesca.Lisi@uniba.it (A. 2)
ORCID: 0000-0002-5287-0468 (A. 1); 0000-0001-5414-5844 (A. 2)

- Create robust and gender-inclusive AI principles, guidelines and codes of ethics within the industry.

The paper is structured as follows. Section 2 addresses the problem of gender bias in ML applications and discusses some notable cases. Section 3 proposes an approach for developing a Gendered Innovation in AI and Section 4 concludes the paper.

2. Gender bias in Machine Learning applications

In this section we wonder about the *fairness* of ML algorithms from a gender point of view. The question is the following: are ML tools, algorithms and technologies, gender neutral? We observe that when following the data-driven paradigm underlying ML, it is necessary to check whether the data used to train the algorithms includes all the *bias* about gender and other possible areas of discrimination - for example ethnicity - diffused in the society. The answer is positive, since for any ML system, the output is determined by the training data, in some cases driven by literally millions of examples. So, these kinds of algorithms – in particular, Neural Networks and Deep Learning – being conceived as learning systems, can upload the gender bias diffused in the society as shown in many examples reported in [5]. The problem arises mainly because little attention is paid to how data is collected, processed and organized. Indeed, the biases are substantially data-driven biases [6].

We cite Joshua Bengio of the Montreal University who told: 'AI can amplify discrimination and biases, such gender or racial discrimination, because those are present in the data the technology is trained on, reflecting people's behavior.' In other words: should we let data speak for itself?

Recently many studies have shown how these ML techniques have brought to applications affected by biases in different fields, from machine translation [7] to assessing geodiversity issues [8], from predictors of crime recidivism [9] to predictors in medicine [10]. It should be mentioned that fairness could be guaranteed, by following two different solution approaches:

1. *Data debiasing* (such as in, e.g., [11])
2. *Model debiasing* (such as in, e.g., [12])

A state-of-the-art survey of works on bias and fairness goes beyond the scope of this paper. The interested reader might refer to, e.g., [13] for an overview of problems and solutions in ML concerning these crucial aspects. For the sake of brevity, and just for illustrative purposes, in the remainder of this section we will focus on three applications which we deem particularly representative: face recognition, word embedding, recruiting tools.

2.1. Face recognition

The systems for face recognition are increasingly used. However, often they turn out to be insufficient for the proper recognition of people of different genders and races. Interesting results about facial recognition technology determined by Joy Buolamwini, a researcher at the M.I.T. Media Lab, have proved how some of the biases in the real world can seep into the facial recognition computer systems [14, 15]. The author has directly experimented that the face of a black person may not be recognized unless wearing a white mask.

The performance of three leading face recognition systems - by Microsoft, IBM and Megvii of China - were studied by classifying how well they could guess not only the gender of an individual but also a man or a woman with different skin tones. The average predictive accuracy percentages obtained were the following ones:

- Lighter male 99 %
- Lighter female 93 %
- Darker male 88 %
- Darker female 65 %

The conclusion drawn by the author was: "A.I. software (we should say M.L. software) is only as

smart as the data used to train it. If there are many more white men than black women in the system, it will be worse at identifying the black women."

The huge amount of data used to train the system has to be balanced respecting gender and racial composition of the population.

In a recent work [16] the problem of face recognition has been addressed by making the system learn demographic information prior to learning the attribute detection task. The system called InclusiveFaceNet, detects face attributes by transferring race and gender representations learned from a held-out dataset of public race and gender identities. With this integration, the approach produces satisfactory results.

In 2020 there has been a policy change for facial recognition:

- IBM quits the facial-recognition business. IBM will no longer sell "general purpose" facial-recognition technology. Reforms and policy proposals have to address racial disparities, the company opposes using technology for mass surveillance, racial profiling and violations of human rights.
- Amazon halts police use of its facial recognition technology.

2.2. Word embedding

The word embedding tools show that the gender bias diffused in the society can be uploaded in the system. In word embedding models, the representation of each word in a high dimensional vector allows to detect the semantic relations between words as words with similar meaning occupy similar parts of the vector space. It has been proved that these tools capture common stereotypes about women and men [17]. In fact, when asking the database "father : doctor :: mother : x", the answer is x=nurse. And the query "man: computer programmer :: woman : x" gives x=homemaker.

The word embedding tools can be terribly sexist as the society is and this constitutes another important example in which a blind application of ML algorithms can lead to a strong reinforcement of existing social and gender biases. As already mentioned, this is due substantially to the mechanisms on which these AI methods are rooted - mainly bottom-up and data-driven methods. It is important to be aware that the algorithms and the tools utilized to solve different problems are not neutral and have to be analyzed deeply in this respect before application.

Aware of this kind of not neutrality, the word embedding biasing capability has been exploited as a quantitative lens to study the evolution of stereotypes and attitudes toward woman and ethnic minorities in the 20th and 21st centuries in the United States [18]. This work provides an approach for temporal analysis of word embedding and shows a new interesting intersection between ML and social science.

As proposed in [17] it is possible to de-bias the database since a vector space is a mathematical object and it can be dealt with mathematical tools. To this aim it is sufficient to clean the database searching for the couples "he : she" that belong to a list of gender biased pairs that need to be removed. In this work the result is a vector space in which the gender bias is significantly reduced. "One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings," say Bolukbasi and co. "However, by reducing the bias in today's computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society." That seems a worthy goal. As the Boston team concludes: "At the very least, machine learning should not be used to inadvertently amplify these biases." The problem is that it is not always possible to clean the database utilizing a list of possible biased gender couples to be compared with the complete list of couples W/M.

2.3. Recruitment

Recently ML specialists at Amazon uncovered a big problem: "their new recruiting engine did not like women" [19]. The company's experimental hiring tool used ML algorithms to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon. The system was trained to vet applicants by observing patterns in resumes submitted to the company over

a 10-year period. Most came from men, a reflection of male dominance across the tech industry. So relentlessly the automatic recruitment tool preferred male candidates. The system was completely changed and, as reported in [16], "Amazon's recruiters looked at the recommendations generated by the tool when searching for new hires, but never relied solely on those rankings, they said.

3. An approach to Gendered Innovations in Science

Gendered Innovations harness the creative power of sex and gender analysis for innovation and discovery. The most prominent researcher in this field, Londa Schiebinger, reports many case studies in different disciplines, starting from the pregnant crash test dummies to machine translation, from heart diseases in women to osteoporosis in men, from assistive technology for the elderly to urban transport plan [20]. Overall, the collection provides a wide roadmap for sex and gender analysis in order to promote reproducible, innovative and responsible research [21].

Considering gender may: (i) add a valuable dimension to research, and (ii) take research in new directions. For instance, research on heart diseases offers one of the most developed examples of gendered innovations. It considers the fact that ischemic heart disease is the leading cause of death for women of US and European populations.

3.1. Gender dimension in Science

Let us see how a new gendered Science can be developed together with new interpretations of facts with respect to a universal male-point-of-view proposed as neutral.

In general, it is important to understand how we can re-design the scientific theories, how we can propose new hypothesis taking into account the gender dimension, how we can formulate new scientific questions having the awareness that another science is possible, how we can produce a critical view of the method in re-shaping the science. According to [22], "There is a need to go beyond stereotypical feminization of products – so called "pinking" – as female preferences can be drivers for substantial innovation", the "pinking" method is not sufficient to produce a new gendered innovation.

Another point to take into consideration is the difference that women and men have in their approach to the use of technology. While women tend to be more interested in the ease of use of technological devices and in their social benefits, many men focus on the performance of the technology and often, technological devices can become for them quite a 'status symbol'. Also, social needs and life models are different for women and men: this can largely influence technology and its products. Since women represent the mentality, the preferences and the needs of every day by more than 50 % of the human race it is important that, as reported in [22]: "If research institutions and industry want to create valuable and sustainable research results and technologies for people (the market), it is recommended to include women at all stages of the research and innovation process".

In [23, 24] we have studied this problem in the field of human-machine interaction showing that the gender dimension influences in an important way the design of robots for assisting and interacting with people. In scenarios where robots can assume complex behaviors, it is very important to consider the gender factor for better results in terms of robot's robustness and efficiency in running the various tasks.

3.2. From confirming to falsifying argument

With these premises, let us now consider a formal reflection on the scientific method and a critical analysis of logical rules underlying the method used in Science [25].

A very common belief is that, in the first instance, experiments are conducted to test the hypothesis of a theory: if the expected observations of experiments are verified then the theory is fully demonstrated. Formally, if the assumptions of the theory are H and O the observations, the rule underlying the knowledge process can be the following:

$$H \rightarrow O \text{ and}$$
$$O$$
$$\text{----------}$$
$$H$$

From the premises that *H* implies *O* and *O* is true, we can deduce that *H* is true. The logical rule that represents this schema goes under the name of *confirming argument*: it seems well representing the process of innovation in scientific research. But it is a wrong logical rule, a fallacy of the sillogism, i.e., an error of the reasoning. It is called the *fallacy of affirmation of the consequent* [26].

It is easy to verify that, given the logical propositions *p* and *q*, the formula:

$$(((p \rightarrow q) \wedge q) \rightarrow p \ )$$

is not a logical tautology.

More in general, as suggested by Popper's theory [27] and Kuhn's thought [28], Science does not proceed for confirming argument and does not advance according to the progressive and continuous accumulation of truth and knowledge.

Science proceeds thanks to the attempts of refutation of the theories proposed. In other words, we advance if there are errors in the accepted theory. So, the right logical rule associated to the production of innovation is called *falsifying argument*, represented by:

$$H \rightarrow O \text{ and}$$
$$\neg O$$
$$\text{----------}$$
$$\neg H$$

From the premises $H \rightarrow O$ (*H* implies *O*) and $\neg O$ (not *O*, *O* false) it can be deduced $\neg H$ (not *H*, *H* false). In other words, when the consequences of a theory are not verified in the experimental context then the theory needs to be completely re-designed. This argument corresponds to the correct logical rule called *Modus Tollens*.

3.3. Gender in Computer Science

The falsifying argument rule can be the basis of a scientific theory that takes gender into account.

Suppose that a certain theory H does not consider the gender dimension (e.g., medicine vs gender medicine). We need to put the following question: following the implication $H \rightarrow O$, do we expect to find the observations *O* foreseen by the theory true *H*?

Evidently not, because 50% of the users of the innovations are women but, as evidenced by a large literature, it is presumable state that the needs of this part of users are not incorporated in the theory for innovation. Hence these observations can be false ($\neg O$) and the theories of departure, too ($\neg H$).

The rule underlying the scientific method in the production of gendered innovations is just the *falsifying argument*. This leads us to say that, in order to produce a new gendered science in all fields, it is not sufficient to apply the 'pinking method' but it is necessary to radically change the assumptions. Only a complete redefinition of the method and the research model with new applications and new ways of observation can re-design the science in a gender perspective. Thus, in order to design AI-based Computer Systems able to socially interact for facing complex challenges, the gender dimension needs to be taken explicitly into account by re-formulating the questions that can produce responsible research innovations.

4. Conclusions

The problem we addressed in this paper is surely very complex. However, it is crucial for the implementation of a Trustworthy AI that developers and users of AI-based tools do not pursue a blind application of data-driven AI methods [29,30].

This is only one of the aspects that assess the vulnerability of ML algorithms to adversarial attacks (both at training and test time). Indeed, the blind application of ML algorithms can lead to a strong reinforcement of existing social and gender bias. So, when we use ML tools we should check whether the data used for training the underlying algorithms includes also all the bias about gender and

ethnicity diffused in the society. In particular, in order to train systems on balanced data sets, it is very important to apply the debiasing method, i.e. a vector space can be cleaned from bias by compiling a list of gender biased pairs to remove this warp. This has been applied in many applications. More in general, new methods for debiasing data should be studied in order to develop Responsible Gendered Research Innovation.

Aware of these problems that affect the fairness of many algorithms, the next step should be to address the problem of how the gender dimension can be taken into account in the content of the scientific production both from a methodological point of view and from the applicative one [23,24, 25]. We have shown on the basis of a formal reflection on the scientific method and a critical analysis of logical rules underlying the method used in Science that new Gendered Science can be developed formulating new scientific questions with the awareness that another science is possible.

5. Acknowledgements

A special thanks to Lorenza Perini of the University of Padova who contributed insights and expertise to this research in the past.

6. References

[1] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, F. Fuso Nerini, The role of Artificial Intelligence in achieving the Sustainable Development Goals. Nature Communications 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y

[2] European Institute for Gender Equality, https://eige.europa.eu/thesaurus/overview

[3] C. Bodei, L. Pagli, L'informatica non è un paese per donne. Mondo Digitale, 2017.

[4] UNESCO. Artificial Intelligence and Gender Equality. Key findings of UNESCO's Global Dialogue (2020). https://en.unesco.org/AI-and-GE-2020

[5] J. Zou, L. Schiebinger. AI can be sexist and racist – it's time to make it fair. Nature 559 (2018): 324-326, https://www.nature.com/articles/d41586-018-05707-8

[6] K. Hammond, 5 unexpected bias in artificial intelligence. TechTrunch. 2016.

[7] G. Stanovsky, N. A. Smith, L. Zettlemoyer, Evaluating Gender Bias in Machine Translation. In A. Korhonen, D. R. Traum, L. Màrquez: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. Association for Computational Linguistics 2019, ISBN 978-1-950737-48-2: 1679-1684, https://www.aclweb.org/anthology/P19-1164/

[8] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, D. Sculley, No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, NIPS 2017 Workshop on Machine Learning for the Developing World, https://arxiv.org/abs/1711.08536

[9] COMPAS, http://www.equivant.com/solutions/inmate-classification

[10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017): 115-118, https://www.nature.com/articles/nature21056

[11] D. Nozza, C. Volpetti, E. Fersini: Unintended Bias in Misogyny Detection. In: P. M. Barnaghi, G. Gottlob, Y. Manolopoulos, T. Tzouramanis, A. Vakali (Eds.): 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, Thessaloniki, Greece, October 14-17, 2019. ACM 2019, ISBN 978-1-4503-6934-3: 149-155

[12] Y. Qian, U. Muaz, B. Zhang, J. Won Hyun: Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. In F. E. Alva-Manchego, E. Choi, D. Khashabi (Eds.): Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop. Association for Computational Linguistics 2019, ISBN 978-1-950737-47-5: 223-228

[13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, Aram Galstyan: A Survey on Bias and Fairness in Machine Learning. CoRR abs/1908.09635 (2019)
[14] S. Lohr, Facial Recognition is Accurate, if You're a White Guy. The New York Times (2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html
[15] J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proc. of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research 81:77-91 (2018), http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
[16] H. Jung Ryu, H. Adam, M. Mitchell, InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity, ICML 2018 Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden.
[17] T. Bolukbasi, K.–W. Chang, J.Y. Zou, V.Saligrama, A.T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems, pp 4349-4357, 2016.
[18] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences of the United States of America, April 17, 2018, 115 (16) https://doi.org/10.1073/pnas.1720347115
[19] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 2018.
[20] L. Schiebinger et al (eds), Gendered innovations in Science, Health&medicine, Engineering and Environment. https://genderedinnovations.stanford.edu
[21] C. Tannenbaum, R.P. Ellis, F. Eyssel, F. et al., Sex and gender analysis improves science and engineering. Nature 575, 137–146 (2019). https://doi.org/10.1038/s41586-019-1657-6
[22] Sanchez de Madariaga. http://www.genderste.eu/i_research01.html , 2013.
[23] S. Badaloni, L. Perini, The influence of the gender dimension in human-robot interaction, in S.M. Anzalone, A. Farinelli, A. Finzi, F. Mastrogiovanni: Proceedings of the 4th Italian Workshop on Artificial Intelligence and Robotics A workshop of the XVI International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, November 14-15, 2017. CEUR Workshop Proceedings 2054, CEUR-WS.org 2018, http://ceur-ws.org/Vol-2054/paper9.pdf
[24] G. Beraldo, S. Di Battista, S. Badaloni, E. Menegatti, M. Pivetti, Sex differences in expectations and perception of a social robot, 2018 IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO 2018, Genova, Italy, September 27-29, 2018. IEEE 2018, ISBN 978-1-5386-8037-7, https://ieeexplore.ieee.org/document/8625826
[25] S. Badaloni, L. Perini, Are algorithms gender neutral? 10th Conf. on Gender Equality in Higher Education, Dublin, 2018. https://genderequalityconference2018.com/
[26] G. Federspil. Logica clinica. I principi del metodo in medicina. Mc-Graw-Hill, Pub. Group Italia, Milano, 2004.
[27] K.R. Popper. The Logic of Scientific Discovery. Routledge Classics, 1959
[28] T. Kuhn. The Structure of Scientific Revolutions (1st ed.). University of Chicago Press, 1962.
[29] High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI. European Commission, Brussels (2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[30] F.A. Lisi, Recent results and activities in Trustworthy Artificial Intelligence. In: Book of abstracts of "The 'Good' Algorithm? – Artificial Intelligence, Ethics, Law, Health", Vatican City, Feb. 26-28, 2020, p. 27 http://www.academyforlife.va/content/pav/it/events/workshop-intelligenza-artificiale.html