Gender knowledge and Artificial Intelligence*

Silvana Badaloni[1,2,†], Antonio Rodà[1,*,†]

[1]Department of Information Engineering, via Gradenigo, 6, 35131 Padova, Italy
[2]Elena Cornaro Center on Gender Studies, University of Padova, Italy

Abstract

Among the various types of biases that can be recognised in the behaviour of algorithms learning from data, gender-related biases assume particular importance in certain contexts, such as the Italian one, traditionally linked to a patriarchal vision of society. This becomes even more true considering the context of university education, where there is a strong under-representation of female students in STEM Faculties, and, particularly, in Computer Science Courses. After a brief review of gender biases reported in Machine Learning-based systems, the experience of the teaching "Gender Knowledge and Ethics in Artificial Intelligence" active since A.Y. 2021-22 at the School of Engineering of the University of Padova is presented.

Keywords

gender bias, gendered innovation, fairness, artificial intelligence, machine learning,

1. Introduction

With the spread of applications that use Machine Learning techniques, increasing attention is having the possible consequences that such applications have on an ethical and social level. AI has certainly been a sector of big challenges but also a sector of questions related to the repercussions on people's rights and freedoms. Since our goal is to develop a trustworthy AI, it is appropriate to face an analysis from the point of view of gender, ethnicity, personal and social development of AI tools, algorithms and technologies. In particular, the Ethics guidelines for a Trustworthy AI of the European Commission[1] list seven key requirements that AI systems should meet in order to be trustworthy: Human agency and oversight, Technical robustness and safety, Privacy and Data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental well-being, Accountability.

Let's consider the gender dimension, that is not explicitly included in the EC guidelines. To frame it in the content of innovation to develop gendered innovations "to harness the creative power of sex, gender, and intersectional analysis for innovation and discovery", as Londa Schiebinger, the most prominent expert in this field, claims [1] [2], it is necessary to radically change the assumptions and to formulate new scientific questions about the discovered innovation [2]. Addressing the elaboration of gendered innovation in the field of AI, we have to

1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022

*Corresponding author.
†These authors contributed equally.
silvana.badaloni@unipd.it (S. Badaloni); antonio.roda@unipd.it (A. Rodà)

[1]https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
[2]http://genderedinnovations.stanford.edu/index.html

take into account that studies in the sub-field of Machine learning have shown that these kinds of algorithms can upload the gender biases diffused in the society [3].

As bias we intend a system of shared knowledge in society, for or against something: biases are based on stereotypes and prejudices. Stereotypes can help us to deal with a different/unknown world. They become central when our world is threatened, they can take the status of judgments that categorize facts or people: in other words, they become prejudices. At this point a vicious circle can arise between stereotypes, prejudices that can lead to discrimination. Machine learning algorithms, like people, are vulnerable to these distortions.

The problem exists since Machine Learning algorithms, by their intrinsic nature, are trained on the basis of training examples, they learn from data, and therefore can subsume and capture the stereotypes related to people sharing a given characteristic, for example the gender identity, which run through the data. If used to make automatic decisions, these potentially biased systems could lead to unfair, incorrect decisions that could discriminate some groups over others. There is the risk of being discriminatory for certain categories of users.

All this information was the stimulus for us to design a Course "Gender Knowledge and Ethics in Artificial Intelligence" that was held for the first time in the academic year 2021/22 in the School of Engineering at the University of Padua to introduce an ethical dimension applied to AI discipline.

In the first part of the paper we will present some relevant case studies were gender biases were found in machine learning-based applications, in order to give a general overview of how the problem can lead to unfair and discriminatory results. Then, in the second part we will describe and discuss our experience of teaching the course as an important way to disseminate a gender culture and to address some of the problems connected to the use of ML-algorithms.

2. Gender bias

While the concept of bias is very broad, gender-related biases are considered an essential aspect of fairness [4]. In particular, we believe that in the Italian socio-cultural context, the gender biases represents a particularly interesting case study for the Artificial Intelligence community, for several reasons listed below.

First of all, numerous studies have shown that gender biases are deeply rooted in our society. Therefore, the risk that the datasets used for many applications with great social impact (autonomous driving vehicles, recommendation systems, personnel selection systems, etc.) contain biases linked directly or indirectly to gender is very high.

Secondly, gender biases affect more or less half of the population, so their presence has an impact on a large number of people.

Thirdly, given the wide spread of gender bias in our societies, it is relatively easy to find datasets on which to experiment with analysis and debiasing techniques.

Fourthly, in comparison with other types of bias (racial, social, etc.), it is easier to define the categories subject to possible discrimination. Gender studies, while recognising the multiplicity of gender identities, validate the existence of two well-defined polarities, male and female. The existence of two prevailing categories facilitates the definition of experimental protocols for the validation of analysis and debiasing techniques.

Fifthly, following the usual practice of bringing our research experiences back into teaching,

promoting studies on gender biases in AI can facilitate the introduction of gender issues into our computer science courses, with a twofold advantage: a) increasing the degree of involvement of our female students, and b) making our male students aware of stereotypes and biases that risk discriminating against their female counterparts, making their university and professional careers more difficult.

3. Biased Machine Learning

Gender bias in AI was reported in various services with an high social impact [5]: employment, medical health, mortage lending, justice systems, and in new applications of technology such as autonomous vehicles. Let's see some relevant cases in different domains.

The outcomes of three commercial gender classification systems (by Microsoft, IBM, and Face++), tested on the Pilot Parliaments Benchmark – a dataset with a balanced intersectional representation on the basis of gender and skin type – showed that all classifiers perform better on male faces than female faces (with a difference in error rate between 8.1% and 20.6%), and on lighter faces than darker faces (difference between 11.8% and 19.2%) [6]. Again with reference to computer vision techniques, gender discrimination was also found in several algorithms for pedestrian detection [7]. This task is particularly sensitive, because disparities in these algorithms could translate into disparate impact in the form of biased accident outcomes. The analysis involved the 24 top-performing methods of the Caltech Pedestrian Detection Benchmark and showed that, on average, children have higher miss rate than adults, and female a higher miss rate than males, tested on the INRIA Person Dataset [8]. However, the problem goes beyond mere computer vision, involving other applications such as automatic recommendation systems.

A field test on the Facebook platform found that an advertisement promoting careers in STEM was showed more times to males than females [9]. In this case, the analysis of the outcomes showed that the bias was coded in the algorithm. Indeed, the system chose to deliver ads more to men than to women because it was designed to optimise ad delivery while keeping costs low. And the cost of an advertisement is higher if it is delivered to a woman than a man, as a consequence of the fact that women are more attractive targets as consumers (indeed, they drive 70% to 80% of all consumer purchases).

Another case was reported by Amazon, that started using a hiring tool to help rank candidates using data from previous hires [10]. The system was shown to systematically downgrade female candidates and, generally, all resumes containing the word women. Interestingly, gender-based discrimination appears to be difficult to prevent due, among other causes, to a history of gender-biased hiring practices that permeate the data. This illustrates how an algorithm can potentially reinforce, and indefinitely perpetuate, already widespread discriminatory practices.

Gender bias is also an open issue for applications based on Natural Language Processing [11]. How word embedding learns stereotypes has been the focus of research on gender bias and artificial intelligence [12]. Since word embeddings are used as a knowledge base in many applications, biases in these models can propagate into many NLP applications. E.g., experiments show in many articles, papers, and websites more female names being tagged as non-person than male names, amplifying gender stereotyping [13]. In general, gender biases diffused in the text used for word-embedding – a condition often verified in textual corpora coming from writings of

the last decades – are subsumed by the model: for example, words related to traditionally male professions are found closer to inherently gendered words, such as *he* or *man*, and vice versa. Techniques to reduce these biases have been recently studied [14], but the problem is not fully solved, in particular for those language that are more grammatically gendered, as Italian [15]. Reducing gender biases in textual corpora is a particularly difficult task also because automatic detection of gender bias beyond the word level requires an understanding of the semantics of written human language, which remains an open problem and successful approaches are restricted to specific domains and tasks.

A recent extensive review [5] identifies eight factors that contribute to gender bias in AI, among them a) the lack of diversity in training data and developers, b) the presence of gender stereotypes in society that are subsumed by the training data, c) programmer bias that consciously or unconsciously also seeps into the algorithm. This review supports the idea that a multidisciplinary approach is needed to address the multiple factors that condition the development of a trustworthy AI.

4. The experience of the Course

In the perspective of developing an Artificial Intelligence you can trust in an inclusive and ethical way, we have taught since A.Y. 2021-22 at the School of Engineering of the University of Padua the course "Gender Knowledge and Ethics in Artificial Intelligence" with the aim to provide the related basic knowledge and principles in a multidisciplinary and interdisciplinary approach. The course (6 CFU, 48 hour of teaching) is opened both to bachelor and master students. The first edition of the course, not compulsory for any course of study, was attended by about 100 students, with a gender distribution in line with that of the engineering school. This means that the course was chosen, with due proportion, by both males and females and the explicit reference in the title to gender knowledge did not alienate male students, as we initially feared. About 80% of the attendees were bachelor's students, testifying to the great interest in the subject even among the youngest students. Despite the fact that we considered these topics more suitable for master's students, the interest shown by the bachelor students made us realise the importance of introducing concepts relating to gender knowledge and ethics even in the first years of engineering, especially considering that many computer engineering students enter the world of work immediately after their bachelor's degree. Most of the participants came from computer engineering and biomedical engineering courses, with some students from electronic engineering.

As concern the contents of the course, the encounter between machines and people in contemporary society raises very central ethical questions. To this aim, it is necessary to introduce an ethical dimension applied to this discipline with a special attention to Machine Learning, facing an analysis from the point of view of gender, ethnicity, personal and social development of the ML algorithms which can lead in some cases to unfair and discriminatory decisions. External experts have been invited to take lectures and seminars, for all the disciplinary fields outside the computer science area. The syllabus and other information on the teaching can be found on the webpage https://www.dei.unipd.it/node/35894.

In a first part of the course, particular attention has been given to some concepts concerning gender equality and gender knowledge in order to contrast stereotypes and prejudices that condition social interactions and to favor a change towards a more equitable and sustainable

society. After an analysis of the differences between sex and gender [16], an attention has been given to gender statistics that characterize the world of the Academy and that have made it possible to draw up the Gender Balance. A critical reflection on the non-neutrality of knowledge and its transmission has been proposed together with an intertwining of gender, science and technology, and the centrality of a gender approach in the field of innovation to develop gendered innovation in different fields of knowledge and, in particular, in the field of Artificial Intelligence

The open ethical questions are many raised by the encounter between machines, intelligent systems and people in the contemporary society. They concern, for example, the definition of a new concept of privacy in face of a meticulous collection of data to which people are subject, the fairness of decisions made by systems based on ML algorithms, the ethics to be adopted for autonomous systems to take decisions in emergency situations, and so on. In our course we have addressed all these problems dealing with ethical and legal issues applied to Artificial Intelligence. These issues and the computational techniques useful to mitigate the possible negative effects, have been discussed in relations to some relevant case studies.

The participation of the students has been wide and lively. In general we can say that this teaching experience was very positive for both teachers and students.

5. Conclusions

As noted by [4], developers of artificial intelligence are overwhelmingly male, whereas those who have reported and are seeking to address this issue are overwhelmingly female (Kate Crawford, Fei-Fei Li and Joy Buolamwini to name but a few). We strongly believe that to mitigate the presence of gender biases in computer science, diversity in the area of machine learning is essential.

It is very important to act on several perspectives. First of all, by reducing the strong under-representation of women in Artificial Intelligence. Advancing women's careers in AI, therefore, is not only a right in itself; it is essential to prevent biases and improve the efficacy of AI-based systems. Then, it is necessary to disseminate a gender culture at different levels, especially toward the younger generation, as the experience of our course has clearly shown. And an updating of training and education programmes in computer science, following a multidisciplinary approach, is perhaps one of the most promising ways to achieve these goals. Moreover, in light of the principles of gendered innovation, we believe that the study of computational techniques to analyse the presence of gender bias and mitigate its effect on outcomes represents not only an interesting problem, but first and foremost a great opportunity.

In the perspective of developing a trustworthy AI able to learn fair AI models even in spite of biased data, it is then important 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 tools.

Acknowledgments

A special thanks to Francesca A. Lisi of the University of Bari "Aldo Moro" for her collaboration to the present research and the lessons held during the Course. This work is partially supported by the project "Creative Recommendations to avoid Unfair Bottlenecks" of the Dept of Information Engineering of the University of Padova.

References

[1] C. Tannenbaum, R. P. Ellis, F. Eyssel, J. Zou, L. Schiebinger, Sex and gender analysis improves science and engineering, Nature 575 (2019) 137–146.

[2] S. Badaloni, F. A. Lisi, Towards a gendered innovation in ai, in: Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020), volume 1613, 2020, p. 0073.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.

[4] S. Leavy, U. C. Dublin, Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning, in: Proc. of the ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering, Gothenburg, Sweden, 2018.

[5] A. Nadeem, B. Abedin, O. Marjanovic, Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies, in: Proc. of the 31st Australian Conference on Information Systems, New Zealand, 2020, p. 12.

[6] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conf. on fairness, accountability and transparency, 2018, pp. 77–91.

[7] M. Brandao, Age and gender bias in pedestrian detection algorithms, in: Workshop on Fairness Accountability Transparency and Ethics in CV at CVPR 2019, 2019.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1, Ieee, 2005, pp. 886–893.

[9] A. Lambrecht, C. Tucker, Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads, Management science 65 (2019) 2966–2981.

[10] J. Dastin, Amazon scraps secret ai recruiting tool that showed bias against women, in: Ethics of Data and Analytics, Auerbach Publications, 2018, pp. 296–299.

[11] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, Z. Berjawi, Gender bias in text: Origin, taxonomy, and implications, in: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, 2021, pp. 34–44.

[12] 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 29 (2016).

[13] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: Proc. of the 57th Annual Meeting of the Ass. for Comput. Linguistics, 2019, pp. 1630–1640.

[14] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, arXiv preprint arXiv:2005.14050 (2020).

[15] D. Biasion, A. Fabris, G. Silvello, G. A. Susto, Gender bias in italian word embeddings., in: CLiC-it, 2020.

[16] A. Viola, Il sesso è (quasi) tutto: Evoluzione, diversità e medicina di genere, Feltrinelli Editore, 2022.