Teaching gender knowledge and ethics in AI to STEM students

Silvana Badaloni[1,†], Carlo Ferrari[1,†] and Antonio Rodà[1,*,†]

[1]Dept. of Information Engineering, University of Padua, via Gradenigo 6b, Padua, Italy

Abstract

Significant disparities in gender equality still persist across European countries. According to the 2024 Gender Equality Index by the European Institute for Gender Equality, the European Union scored an average of 71.0 out of 100, with wide gaps among member states. The average gender pay gap remains around 12%, and women continue to be underrepresented in STEM fields and leadership positions. These inequalities are also evident in the field of Artificial Intelligence, where in the EU and UK, only 16% of individuals with skills in this field are women. In this context, the article presents the course Gender Knowledge and Ethics in Artificial Intelligence, offered by the School of Engineering at the University of Padua. This initiative, promoted by two teachers of the degree program in Computer Engineering, marks the first explicit introduction of gender-related topics within an engineering curriculum. As the aim of the course is to raise awareness among future graduated about the intersection of gender, ethics, and intelligent technologies, fostering a more inclusive and responsible technical culture, the course was then opened also to STEM students and more generally to all students of the University of Padua. Through both qualitative and quantitative analysis, the article outlines the motivations behind the development of the course, its educational objectives, and its main topics, which include algorithmic bias, fairness, and accountability in AI development. Data collected from the initial editions of the course show consistently high levels of student appreciation and engagement, confirming the course's effectiveness in encouraging critical thinking and promoting a more ethical and inclusive approach to artificial intelligence engineering.

Keywords

Gender knowledge, Artificial Intelligence, Bias, Education in AI,

1. Introduction

In educating new generations of students about the foundations of current AI applications, it is appropriate to face an analysis from the point of view of gender, ethnicity and social status and, more generally all ethics aspects, with respect to personal and social development of AI tools, algorithms and technologies, in line with the vision of trustworthy AI defined by the European Union [1].

Studies in the field of Machine Learning have shown that these kinds of algorithms can incorporate or perpetuate many different types of biases prevalent in society, generating outputs and decisions that can harm historically disadvantaged groups of users. While the concept of bias is very broad, gender-related biases are considered an essential aspect of fairness. In particular, we believe that in the European socio-cultural context, the gender bias 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 widespread use of this type of bias, 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

2nd Workshop on Education for Artificial Intelligence (edu4AI 2025 https://edu4ai.di.unito.it/), co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025). October 26, 2025 in Bologna, Italy

[*]Corresponding author.

[†]These authors contributed equally.

silvana.badaloni@unipd.it (S. Badaloni); carlo.ferrari@unipd.it (C. Ferrari); antonio.roda@unipd.it (A. Rodà)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

CEUR-WS.org/Vol-4114/11_paper.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-01

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 bias in AI and education in AI courses 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 biases that risk discriminating against their female counterparts, making their university and professional careers more difficult.

Furthermore, the under-representation of women in the digital technology sector, particularly in AI, is a significant concern that impacts the fairness and inclusivity of AI frameworks. Data from Europe, the UK, and Italy consistently show that women comprise only 16% of the AI workforce, with an even smaller percentage (12%) having over a decade of experience [1].

All these reflections highlight a complete lack in the educational paths in Engineering and in the STEM areas expecially those devoted to form strong and inclusive AI experts. At the design level it is possible to introduce interdiscipinary courses spanning different topics for a limited number of credits without reducing base and characterizing credits. Then we believe that Artificial Intelligence is a suitable field for introducing ethics and gender-related topics within the various degree in the School of Engineering.

Since 2021, the School of Engineering at the University of Padua has been pioneering an innovative course titled "Gender Knowledge and Ethics in Artificial Intelligence". This initiative, promoted by two teachers of the degree program in Computer Engineering, marks the first explicit introduction of gender-related topics within an engineering curriculum. They specifically started in the bachelor's degree program in Computer Engineering and, given the students outcome, they opened the course to the other STEM students in the following years. Finally the course is nowadays a "General Course" for all students at the University of Padova. The course consists of 48 hours of in-person lectures and has been designed without requiring specific disciplinary prerequisites, allowing it to be open to students from various academic backgrounds. By inviting various experts to deliver lectures and seminars, the course provides a rich, diverse perspective on critical topics such as diversity, equity, and inclusion, particularly noteworthy given the persistent gender disparity within the engineering field.

2. Educational objectives and methodology

The course's fundamental premise is the recognition that technology is not neutral but profoundly shaped by the values, experiences, and perspectives of its creators. This approach challenges the long-held notion of technological neutrality and encourages students to consider the broader societal implications of their work. By integrating these themes into technical curricula, the course offers multiple benefits to students and the field of engineering as a whole. Firstly, it raises awareness among future engineers about the social impact of the technologies they develop. This awareness is crucial in an era where artificial intelligence and other advanced technologies are increasingly influencing various aspects of society. Secondly, the course fosters a more reflective and critical approach to technological development. By encouraging students to question assumptions and consider ethical implications, it helps create more responsible and thoughtful engineers. Moreover, the consideration of diverse perspectives can lead to more innovative and inclusive technological solutions. By exposing students to a variety of viewpoints and experiences, particularly those related to gender, the course helps broaden their understanding and approach to problem-solving. This diversity of thought is essential in creating technologies that serve and represent all members of society. Perhaps most importantly, the course plays a role in shaping a new generation of technology professionals. These future leaders in the field will possess not only technical expertise but also a sense of ethical responsibility and awareness of gender issues. This includes an understanding of stereotypes and biases that characterize both our societies and the machines that have learned from these societal models. By recognizing these biases, students are better equipped to prevent their incorporation into AI systems, leading to more equitable and fair technological solutions.

The intended learning outcomes for this course encompass a comprehensive understanding of AI fundamentals and their ethical implications. In particular, these are:

- Describe the fundamentals of AI and differentiate between symbolic approaches and ML. (Knowledge)
- Explain key ethical principles related to AI. (Knowledge)
- Recognize gender and ethnic biases in algorithms. (Judgement)
- Evaluate ethical impacts of AI in real-world cases. (Applying)
- Propose mitigation strategies. (Applying)
- Collaborate in interdisciplinary teams and communicate analyses clearly. (Communication + Learning)

Topics not belonging to the computer science/engineering area were presented through the intervention of professors from various disciplinary fields, who were invited to deliver one or two lectures to the course students. The computer science/engineering topics, on the other hand, were covered by the main course instructors, with an interdisciplinary approach that takes into account the knowledge acquired in other fields. Regarding the professors who accepted the invitation, they come from psychology, biology, philosophy, law, linguistics, as well as computer science experts in interdisciplinary topics.

During the course, students are organized into groups of 3 to 5 members, and each group is assigned a topic to explore through reading articles or books. Since some students come from various disciplinary areas outside of engineering, the groups that form are often multidisciplinary, which facilitates the development of multiple perspectives on the same topic. In the final lessons of the course, each group presents the results of their study to the other students. The presentation and a short essay written by each group are evaluated and taken into account for the final exam.

3. Main topics

The topics of the course are organized into three pillars: gender knowledge, ethics, and fairness. Whereas ethics and fairness are aspects that are becoming common in AI courses and book, the accent on gender knowledge is quite typical of our course (the reasons of this design choice are explained in details above in Section 1).

3.1. Gender knowledge

Merely increasing the number of women in STEM fields, while important, is insufficient for achieving true inclusivity and fairness. The more profound change required is "fixing the knowledge" by integrating the gender dimension into scientific content, leading to "gendered innovations." This means ensuring that scientific research and development consider both biological (sex) and sociocultural (gender) characteristics, behaviors, and needs of all individuals, without disparities.

The leading expert in gendered innovations is Londa Schiebinger of the Stanford University, who advocates for fundamentally rethinking existing assumptions and formulating new scientific questions to harness the creative power of sex, gender, and intersectional analysis for innovation and discovery [2].

The critical questions posed are: How can a new gendered science be developed, along with new interpretations of facts, in contrast to the traditionally perceived "universal male" point of view in STEM [3]? This points to the need for a paradigm shift, moving away from a historically male-centric perspective that has often been presented as neutral and universally applicable, but which in reality may overlook or misrepresent the experiences and characteristics of women and other marginalized groups. How can we formulate new scientific questions with the awareness that another science is possible? This question encourages a proactive and imaginative approach to scientific inquiry, challenging researchers to envision and pursue lines of questioning that explicitly incorporate gender and other social dimensions from the outset, rather than as an afterthought. How can we create a critical view of the methods used to reshape science? This calls for a metacognitive approach to scientific methodology, prompting researchers to critically examine the assumptions, biases, and limitations inherent in current research methods and to develop new, more inclusive methodologies that can better account for the complexities of sex, gender, and intersectionality.

In essence, a transformative shift in scientific and education practice, moving to fundamentally alter the content and methodology of scientific knowledge creation to be truly inclusive and produce more excellent, relevant, and equitable research outcomes.

3.2. Ethics in AI

In addressing the ethical challenges posed by artificial intelligence, it is crucial to consider the concept of ethical pluralism. This approach acknowledges that there are multiple, sometimes conflicting, ethical frameworks that can be applied to complex moral dilemmas. A poignant example of this, as discussed by Guglielmo Tamburrini [4] in his work on the ethics of autonomous vehicles, is the scenario of an inevitable collision between a self-driving car and one of two bicycles, one ridden by a woman without a helmet and one ridden by a man wearing a helmet. Students are asked the question: Given that a collision is inevitable, which bicycle should the autonomous vehicle's algorithm be programmed to hit? This situation presents a stark ethical dilemma that highlights the divergence between consequentialist and deontological ethical approaches. From a consequentialist perspective, one might argue for minimizing harm by choosing the action that results in the least overall damage or injury, for example by hitting the cyclist wearing the helmet. Conversely, a deontological approach might prioritize the responsible behavior of the person wearing the helmet, thus hitting the person without the helmet. Also because if wearing a helmet were not rewarded, in the long run this would discourage people from following this good rule of behavior, resulting in more injuries.

This type of reasoning can be applied to many different contexts where AI is used, such as in the case of security: is it better to prioritize people's safety by using cameras and facial recognition systems, or to protect their privacy? In this way, students are progressively introduced to the complexity of a world that is increasingly influenced by decisions made by AI-based algorithms.

3.3. Fairness, equity, and mitigation approaches

For AI-based tools to be trustworthy, one of the widely recognized requirements is that the outputs (generated content and decisions) are fair. Due to the intrinsic nature of the machine learning approach, these systems can capture and reinforce the biases present in the society, which are reflected in the datasets used for training. If used to make automated decisions, these systems can lead to "unfair" outcomes that may discriminate against certain groups [5].

While there is agreement on this requirement, it is not as simple to define the concept of fairness [6, 7]. What is considered fair in one cultural context may not be so in others. Moreover, like all socio-cultural constructs, it changes over time and needs to be continuously questioned. Furthermore, even assuming a shared definition of fairness, evaluating the fairness of a computer system requires quantification that is not immediately obtainable: in recent years, dozens of fairness metrics have been defined [8], which are not easy to apply and understand. It is therefore necessary to provide students with a problematic view of the concept of fairness, equipping them with critical and analytical tools to apply both quantitative and qualitative approaches in real-world cases, taking into account the socio-cultural context.

The topic is addressed in class by presenting several case studies, some simple and others more complex. Gender discrimination is used as a common thread: as explained earlier, this allows for greater engagement, given that relationships and conflicts between genders concern everyone and are highly relevant to our students' age group, and it facilitates an awareness of gender disparities present in our societies. Among the simpler and well known cases, we present that of Joy Buolamwini, afro-american researcher at MIT Media Lab. She discovered that the camera system installed in her laboratory did not recognize her well, but when she put on a white mask the system functioned perfectly [9, 10]. She realized that the system's accuracy was systematically higher for white men and lower for black women. Machine learning systems are 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 poorer at identifying the black women.

A more complex case is that of the COMPAS system (Correctional Offender Management Profiling for Alternative Sanctions) [11], a software application used in some counties in the United States to assign a risk score to individuals on trial for committed crimes. The score assigned by the application can be used by the judge to assign alternatives to imprisonment, in case the defendant is deemed not socially dangerous. The system has been accused of being discriminatory against African-American defendants [12]. Indeed, some metrics commonly used to quantify fairness support this accusation. Further analyses, however, have challenged this conclusion, still following a quantitative approach, but using other metrics deemed more suitable to the context [13, 14]. In particular, while the accusation of discrimination against African-Americans is difficult to demonstrate quantitatively, a gender-sensitive approach instead shows unfavorable treatment of the system towards women [14].

4. Evaluation

The fourth edition of the course has recently concluded. Despite being an elective course, the number of participants has remained fairly stable, around 90 students per year, with 35% female ($n = 130$) and 65% male ($n = 239$). This distribution is in line with that of the STEM degrees at the University of Padova (about 37% of females). 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. 70% of the participants come from Computer Engineering and Biomedical Engineering, equally distributed, 10% from other engineering courses, while the remaining 20% come from courses in other disciplinary fields including: data science, law, political science, philosophy, neuroscience, natural and environmental sciences, psychology, linguistics, communication. The final grade does not show significant differences between females ($M = 28.4 SD = 2.2$) and males ($M = 28.2 SD = 2.1$). At the end of the course, students were given a questionnaire to express their opinion[2]: to the question "Overall, how satisfied are you with how the course was conducted?" they gave an average score of 8.34 out of 10 (median 9), which is higher than the average value for computer engineering courses.

Additionally, students were asked to provide an open comment about the course. Below, we have selected a few that we believe well summarize the strengths and weaknesses of the course.

"It’s an important subject for those studying AI. Reflecting on issues we normally don’t consider is essential to create a class of engineers who are conscious of what they’re doing."

"In my opinion, this course is very interesting because it deals with sensitive and important topics that are neglected in all other courses of the degree program."

"A truly interdisciplinary approach that aims to connect very different students and departments on a topic that is perhaps one of the most important of the century."

"The topics covered are numerous, and being so many, they are sometimes, understandably, not explored in depth."

"Being a course designed to have people from different faculties, there’s a tendency to have topics that are very simple for some and very difficult for others: for those who have studied humanities, all the topics regarding ethics are easier to understand and study, but those requiring mathematical calculations are almost impossible."

The course is widely regarded as important and interesting, particularly for STEM students, as it addresses critical ethical issues often neglected in other parts of the curriculum. Students appreciate its interdisciplinary approach, which connects diverse fields and perspectives on what they consider one of the century’s most significant topics. The course is seen as essential in developing conscientious engineers aware of the broader implications of their work. However, some challenges were noted. The wide range of topics covered means that some areas are not explored in great depth. Additionally, the diverse academic backgrounds of students lead to varying levels of difficulty across topics, with humanities students finding ethical discussions more accessible and technical aspects more challenging.

5. Conclusions

The initiative has thus far garnered significant interest from both male and female students, with active participation in lectures demonstrating its relevance and appeal. However, integrating these themes

[2] See https://www.unipd.it/opinione-studenti-sulle-attivita-didattiche for more details about the survey methodology

into technical courses is not without challenges. It may face resistance from faculty accustomed to a purely technical approach. In particular, social and philosophical issues are still considered by some collegues less relevant for the training of engineers and are downgraded to soft skills that are not strictly necessary. As a consequence, the educational committees of some degree programs have refused to include this course in the study plans of some students. These episodes reflect the need for a significant shift in the academic and professional culture of the technology sector.

The 6-credit course we have implemented in the bachelor's degree in Computer Engineering has been well-received by students and provides an introduction to "horizontal" issues in AI ethics. While this course offers a valuable foundation, addressing the professional challenges in this field requires a higher level of specialization. To meet this need, we are currently designing a more advanced course for the master's degree program. This upcoming course will focus on specific techniques for risk analysis and mitigation in AI systems. By offering this additional, more specialized training at the graduate level, we aim to equip our students with the in-depth knowledge and practical skills necessary to tackle the complex ethical challenges they will encounter in their professional careers.

Declaration on Generative AI

During the preparation of this work, the authors used Claude 3.5 for linguistic tasks such as translation and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Thil, D. Barbieri, J. Caisl, G. Lanfredi, J. Linkeviciute, B. Mollard, J. Ochmann, V. Peciukonis, J. Reingarde, M. Kullman, et al., Artificial intelligence, platform work and gender equality (2022).

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

[3] 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.

[4] F. Fossa, G. Tamburrini, Ethics of autonomous vehicles. from moral dilemmas to social trade-offs, Paradigmi 40 (2022) 79–94.

[5] S. Badaloni, A. Rodà, et al., Gender knowledge and artificial intelligence, in: Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA, 2022.

[6] J. Broome, Fairness, Proceedings of the Aristotelian Society 91 (1991) 87–101.

[7] B. Hooker, Fairness, Ethical theory and moral practice 8 (2005) 329–352.

[8] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, A. C. Cosentini, A clarification of the nuances in the fairness metrics landscape, Scientific reports 12 (2022) 4209.

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

[10] S. Lohr, Facial recognition is accurate, if you're a white guy, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 143–147.

[11] T. Brennan, W. Dieterich, B. Ehret, Evaluating the predictive validity of the compas risk and needs assessment system, Criminal Justice and behavior 36 (2009) 21–40.

[12] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of data and analytics, Auerbach Publications, 2022, pp. 254–264.

[13] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, Northpointe Inc 7 (2016) 1–36.

[14] A. Rodà, The COMPAS case: aneducational journey for explaining fairness in ai-based applications, in: Proc. of AIMMES 2025 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies, 2025.