

# Ethics and Gender for Responsible Research and Innovation in AI

Francesca Alessandra Lisi<sup>1,\*</sup>

<sup>1</sup>*Dipartimento di Informatica, University of Bari "Aldo Moro" - Via E. Orabona 4, Bari, 70125, Italy*

## Abstract

This short paper is an extended abstract of the invited talk I gave at the BEWARE 2022 workshop. It addresses two themes that are common to the Responsible Research and Innovation approach, and to the EU guidelines for a Trustworthy Artificial Intelligence: Ethics and Gender. The talk (and this paper) was intended to account for some recent research and dissemination activities concerning these themes, to explore the interplay between the two, and to lay new foundations for AI Ethics inspired by contemporary feminist theories.

## Keywords

AI Ethics, Gender Studies, Responsible Research and Innovation, Trustworthy AI, Feminism

## 1. The need for a responsible approach to AI

Artificial Intelligence (AI) is the latest technological evolution which is transforming the global economy, is playing a major role in the "Fourth Industrial Revolution", and is deeply changing our society. As such, it brings so-called *disruptive technologies*, namely those technologies that have the potential to supersede an older process, product, or habit. In order to reveal how these technologies can be socially and economically transformative, and to identify the potential downsides of the 'disruption' it might bring about, it is necessary to take a responsible approach to them. *Responsible Research and Innovation* (RRI) is just the approach, promoted by European Commission's Science in Society programme in the context of the Horizon 2020 Strategy, that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation. RRI involves 5 thematic elements: Ethics, Gender, Public Engagement, Education, Open Access. These concern holding research to high ethical standards, ensuring gender equality in the scientific community, investing policy-makers with the responsibility to avoid harmful effects of innovation, engaging the communities affected by innovation and ensuring that they have the knowledge necessary to understand the implications by furthering science education and the accessibility of scientific knowledge.

---

*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, affiliated also to *Centro Interdipartimentale di Studi sulle Culture di Genere (CISCuG)* of the University of Bari "Aldo Moro".


✉ [FrancescaAlessandra.Lisi@uniba.it](mailto:FrancescaAlessandra.Lisi@uniba.it) (F. A. Lisi)

🌐 <http://www.di.uniba.it/~lisi/> (F. A. Lisi)

🆔 0000-0001-5414-5844 (F. A. Lisi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

The disruptive technologies brought by AI, although successfully applied in a growing number of contexts, have turned out to suffer from a major problem: The lack of trust from adopters, due to several flaws in the way these technologies are designed, developed and applied in risky domains. As humans, we tend to trust more white-box systems than the black-box ones. So, applications of machine learning/deep learning algorithms are the main source of concern. At the EU level, the problem has been faced by the High-Level Expert Group on Artificial Intelligence of the European Commission in the Ethics Guidelines for a Trustworthy AI [1]. According to the Guidelines, AI deserves trust if it is: (1) *lawful*, i.e. compliant with all applicable laws and regulations; (2) *ethical*, i.e. not violating ethical principles and values; (3) *robust*, from both a technical and social perspective. Diversity, non-discrimination and fairness are among the requirements listed in the Guidelines. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle. In particular, algorithmic biases must be avoided, as they could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination, e.g. based on gender or race.

The themes of Ethics and Gender are therefore common to RRI and Trustworthy AI. Indeed, these are the subject of a wide variety of activities carried out by the AI community in the last years. I will briefly describe some of them in the rest of the paper. The activities considered here span from research to public engagement, since bridging the gap between science and society is the ultimate goal of RRI. A particular emphasis is given to initiatives promoted by the *Italian Association for Artificial Intelligence (AIxIA)*, since the BEWARE 2022 workshop has been held as part of the programme of the 2022 edition of the AIxIA conference.

## 2. AI and Ethics

The debate around AI Ethics is inherently multidisciplinary and interdisciplinary. It often includes the viewpoint of Law experts, as done in a couple of initiatives born within the EU-funded COST Action “DIGital FOREnsics: evidence Analysis via intelligent Systems and Practices” (DigForASP)<sup>1</sup> during 2021. The former event, entitled “DigForEthics: Conversations on Digital Forensics, Ethics, Law and Artificial Intelligence”, was organized by Viviana Mascardi from the University of Genova, Italy, and Juan Carlos Nieves from Umeå University, Sweden.<sup>2</sup> Besides invited talks by prominent researchers like Virginia Dignum and Carles Sierra, it featured roundtables with the aim of discussing the relevance of the seven requirements for a Trustworthy AI. The latter event, organized by Chiara Gallese Nobile from Eindhoven University of Technology, The Netherlands (now in Trieste, Italy), David Billard from the University of Applied Sciences in Geneva, Switzerland, Elena Falletti from Carlo Cattaneo University LIUC, Italy, and myself, was devoted to “Automated Decision Making, AI systems and individual rights”.<sup>3</sup> The conference analyzed the ethical, legal, and technical issues of those systems used to take a decision by automated means without any human intervention.

---

<sup>1</sup>DigForASP (<https://digforasp.uca.es/>) explores the potential of the application of AI (in particular, Automated Reasoning) in the Digital Forensics field. The Action started in Sept. 2018 and will end in March 2023.

<sup>2</sup>For a brief report see: <https://www.cost.eu/when-artificial-intelligence-meets-digital-forensics-ethics-and-law/>

<sup>3</sup>Report here: <https://www.cost.eu/artificial-intelligence-systems-and-individual-rights-where-are-we-heading-to/>

AI Ethics is not just a subject for exciting debates across disciplines. It should also be of practical interest for AI researchers and developers, since ethical issues are raised by many AI applications. Chatbots are among the most critical ones. These are tools aimed at simplifying the interaction between humans and computers, typically used in dialogue systems for various practical purposes, *e.g.*, in customer service. Their behavior may heavily influence a user and may reflect some bias, *e.g.*, against women. It is therefore important that these systems are built upon ethical foundations. Drawing from practical philosophy, it can be argued that AI-based systems could develop ethical decision-making and judgment capabilities by learning from experience. This has been inspiration for the work reported in [2] that addresses the problem of evaluating the ethical behaviour of AI-based chatbots in an online customer service chat point w.r.t their institution/company's codes of ethics and conduct. Here, the proposed approach combines two logic-based AI techniques, Answer Set Programming (ASP) and Inductive Logic Programming (ILP), for defining the detailed ethical rules that cover real-world situations from interactions with customers over time [3, 4]. ASP is appropriate for representing and reasoning with ethical rules because it can deal with norms and exceptions, whereas ILP can automatically generate those ethical rules that are difficult to encode manually. The approach is implemented in the form of modules for a Multi-Agent System (MAS), which includes a component for text extraction from dialogues [5].

A major problem in AI Ethics is the current lack of benchmarks for evaluating and assessing the adherence of AI systems to the ethical standards and the norms of their environment, as stressed also during the discussion at the AAAI Spring Symposium 2021 on implementing AI Ethics.<sup>4</sup> A collection of unethical scenarios, with reference to different application domains, would help a lot. In order to address this open issue, the MAS prototype presented in [5] has a great potential for future implementations of ethical chatbots in different domains. Indeed, simulation by means of a MAS might provide better insights into the dynamics of a corresponding real-world system, and to assess the practical challenges and limitations of building such a system. However, the learning module in the MAS needs data to train the proposed customer service chatbot. This is still a major weakness in technical solutions for AI Ethics which leverage some machine learning algorithm.

### 3. AI and Gender

The relationship between AI and Gender is particularly strained. Several documents delivered by international organizations like UNESCO, see [6, 7, 8], over the past few years have stressed the risks of AI for Gender Equality. In particular, it has been shown that current AI-based technologies can not be considered an enabling factor for reaching the fifth<sup>5</sup> of the Sustainable Development Goals (SDGs) in the Agenda 2030 launched by United Nations in 2015 [9].<sup>6</sup> Thus, the Gender theme deserves the greatest attention. An increasing number of initiatives have been recently undertaken, some more focused, other devoted to the broader theme of Diversity

---

<sup>4</sup><https://aaai.org/Symposia/Spring/sss21symposia.php#ss06>

<sup>5</sup>SDG 5: Gender Equality and Women Empowerment

<sup>6</sup>See also the book "L'Intelligenza Artificiale per lo Sviluppo Sostenibile" freely available at <https://www.cnr.it/sites/default/files/public/media/attivita/editoria/VOLUMEFULL14digitalLIGHT.pdf>

& Inclusion, inside or outside the AI community. Among scientific organizations in the AI field, the *Association for Advances in Artificial Intelligence* (AAAI) shows to be the most sensitive to these aspects and most active in organizing events for raising awareness of them in the scientific community. Since 2020, the annual AAAI conference hosts an entire programme of workshops on Diversity and Inclusion. At the national level, starting from 2020 as well, AIxIA has organized several events, among which the online conferences “Questioni di genere in Intelligenza Artificiale”<sup>7</sup> and “Intelligenza Artificiale e impatti di genere: una questione urgente”<sup>8</sup> were targeted to the general public and to the corporate management, respectively. The former, organized and chaired by myself, took place on November 25, 2020, as part of the annual conference of AIxIA, and was the first Italian event of the kind. In the latter, held on July 13, 2021, I was invited to give a talk on the role of Diversity & Inclusion principles to implement the vision of Trustworthy AI. The challenges of Diversity in AI were also discussed in my talk at a conference on digital technologies, AI and gender issues organized by Ines Crispini from the University of Calabria on March 7, 2022, under the patronage of AIxIA, together with the *Società Italiana per l’Etica dell’Intelligenza Artificiale* (SIpEIA).<sup>9</sup>

It has been already mentioned that AI, although it can act as a catalyst to achieve the goals of the Agenda 2030, it may also trigger inequalities that turn out to act as inhibitors on SDG 5 (and related SDGs). There are a few projects that are trying to counter-balance the many shortcomings of AI-based technologies with respect to the gender dimension, by developing applications that can effectively support the achievement of SDG 5. Among them, there is the proposal of using AI techniques for studying *gender power relations*, namely the ways in which gender shapes the distributions of power at all levels of society. These relations are among the most persistent patterns in the distribution of power, and strengthen inequalities between women and men. The set of roles, behaviours and attitudes that societies define as appropriate for women and men (‘gender’) can be the cause, consequence and mechanism of power relations, from the intimate sphere of the household to the highest levels of political decision-making. Wider structures and institutions can also shape the distribution of power by reinforcing and relying on gender roles. In the project, funded by the University of Bari in 2022,<sup>10</sup> institutional documents have been analyzed with NLP algorithms in order to detect these patterns at the linguistic level. The basic idea is that the discovered patterns can be further elaborated with logic-based AI techniques such as ILP to derive policy rules. In this project I am the Key Area Person for the team of computer scientists and mathematicians, which includes the NLP researchers Marco de Gemmis and Pierpaolo Basile among the others.

The relationship between AI and Gender can be read also in the opposite direction: What can Gender Studies do for AI? Particularly interesting and challenging is the promotion of a so-called *Gendered Innovation* in AI. Gendered Innovations harness the creative power of sex, gender, and intersectional analysis for innovation and discovery.<sup>11</sup> The adoption of these approaches may add valuable dimensions to research, as shown for many disciplines ranging

---

<sup>7</sup><https://aixia2020.di.unito.it/program/questioni-genere-ia>

<sup>8</sup><https://aixia2021.disco.unimib.it/satellite-events-ita>

<sup>9</sup><https://sipeia.diag.uniroma1.it/wp-content/uploads/2022/02/Ines-Crispini.pdf>

<sup>10</sup>[https://www.uniba.it/it/ricerca/centri-interdipartimentali/culture-di-genere/portlet-general/HESeeds\\_abstract\\_eng1.pdf](https://www.uniba.it/it/ricerca/centri-interdipartimentali/culture-di-genere/portlet-general/HESeeds_abstract_eng1.pdf)

<sup>11</sup><http://genderedinnovations.stanford.edu/index.html>

from Medicine to Engineering. In the case of AI, this means, *e.g.*, that the awareness of the problems that affect the fairness of many algorithms, notably ML algorithms, should push towards truly innovative ways of doing research in the field. In particular, 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 from a methodological as well as an application point of view. In [10], starting from a critical analysis of logical rules underlying the scientific method, it is shown that a gendered science can be developed if we start formulating new scientific questions with the awareness that another science is possible. So, it is not only the case that AI brings disruptive technologies but also the case that other disciplines (*e.g.*, Philosophy) may act as an innovation driver for AI, and help overcome some limits of current AI research and practice.

#### 4. Feminism-inspired foundations for AI Ethics

An interesting direction of research for a RRI approach to AI is the one that explores the interplay between the two themes of Ethics and Gender. This should be founded upon an interdisciplinary debate which involves AI researchers/practitioners and scholars from other disciplines, notably Philosophy as already mentioned. In particular, it is noteworthy that recently there has been an increasing interest in *contemporary feminist theories* to face some open issues in AI Ethics, notably as the starting point for solutions aimed at addressing the challenges of diversity, non-discrimination and fairness in Trustworthy AI.

The term Feminism denotes a range of socio-political movements and ideologies that aim to define and establish the political, economic, personal, and social equality of the sexes. It has provided Western women with increased educational opportunities, the right to vote, protections against workplace discrimination, and the right to make personal decisions about pregnancy. In some communities, Feminism has also succeeded in challenging pervasive cultural norms about women. Since the late 20th century, the scope of Feminism has become broader and broader, with the creation of ethnically specific or multicultural forms of feminism, such as *Black Feminism* and *Intersectional Feminism*. These recent evolutions of Gender Studies are having an increasing impact in AI so that they can be considered as an inspiration for new foundations of AI Ethics. In the following I will mention some notable examples of this trend.

First, I would like to mention D.A.K.I.N.I., a multidisciplinary and transdisciplinary performance that aims at investigating and creating dialectic bridges between the field of AI and contemporary feminist theories (in particular, Cyberfeminism and Posthuman Feminism).<sup>12</sup> It stems from an idea of the collective AjaRiot of performing arts in co-production with the Nordisk Teaterlaboratorium (NTL – Odin Teatret, DK) and was internationally developed in 2018 and 2019, also by seeking advice from several AI experts, myself included as a AIxIA delegate. In D.A.K.I.N.I. there are several citations of Donna Haraway's *Cyborg Manifesto* [11], *e.g.*, the awareness that in a world ridden by the *informatics of domination*, women have to deal with the question of their involvement with technology, and face its complexity. The questions to answer are therefore: How can AI serve feminist theories, and life? How can women take control of it? What happens after they do so? What do women want to pass on to future generations? Through a mix of artistic languages (physical theatre and dance-

---

<sup>12</sup><https://ajariotcollective.com/d-a-k-i-n-i/>

theatre, video projections, photographs, interviews, merging of ancient and electronic music, sound experiments) the performance touches several crucial aspects of the relationship between women and technology: (i) the under-representation of women in ICT, both in education and in professions, (ii) their de-facto exclusion from the processes of design, development and test of digital technologies, (iii) the gender biases in machine learning algorithms and automated decision making systems. The ultimate message of the performance is the importance of having role models, here represented as a kind of *dea-ex-machina* (the goddess Dakini, symbol of femininity and women empowerment in Hinduism and Buddhism), to encourage women to take active part in the digital revolution. The epilogue of the performance reflects the thought of Rosi Braidotti [12], *e.g.*, her claim: “I see the *posthuman* turn as an amazing opportunity to decide together what and who we are capable of becoming, and a unique opportunity for humanity to reinvent itself affirmatively, through *creativity* and empowering *ethical relations*, and not only negatively, through vulnerability and fear”.

Leaving the Arts and going back to more technical subjects, the work of Catherine D’Ignazio and Lauren Klein on so-called *Data Feminism* presents a new way of thinking about data science (and data-driven AI) that is informed by the ideas of intersectional analysis [13]. Data Feminism goes far beyond gender. It deals with power, by analyzing who has it and who doesn’t, and how those differentials of power can be challenged and changed. This awareness of the existence of gender-based power relations (as in the project mentioned in the previous Section) raises several questions that the data scientist should answer during his/her professional practice: Data science by whom? Data science for whom? Data science with whose interests in mind? In their book, D’Ignazio and Klein show how challenges to the traditional male/female binary can help detect and discard other hierarchical (and empirically wrong) classification systems, such as the ones resulting from data science for harm (*e.g.*, to discriminate, police, and surveil).

Finally, particularly interesting is the position of Virginia Dignum in [14]. She starts with a general discussion of current conceptualisations of AI followed by an overview of existing approaches to governance and responsible development and use of AI. Then, she reflects over what should be the bases of a social paradigm for AI and how this should be embedded in relational, feminist and non-Western philosophies, in particular the Ubuntu philosophy. Notably, her proposal does take into account the requirement of robustness, as typically done in socio-technical systems and in line with requirements for a Trustworthy AI.

## 5. Conclusions

A responsible approach to research and innovation in AI calls for multidisciplinary and interdisciplinarity. In particular, the relationship with Law and Philosophy can be two-way. On the long run, due to the disruption brought by AI, the interlinking between these and other fields will become tighter and tighter. It will be unavoidable to overcome the current separation between Humanities and Sciences and Engineering. Indeed, Education - one of the other RRI pillars besides Ethics and Gender - plays a prominent role in this transformation, as also testified by some pilot experiments in academic curricula such as the class of “Gender Knowledge and Ethics in Artificial Intelligence” at the School of Engineering of the University of Padua (see [15] for a report of the experience done by the teachers Silvana Badaloni and Antonio Rodà).



## Acknowledgments

This paper is partially based upon work from the Italian project “Future AI Research (FAIR)”, funded by PNRR. The work contributes to the goals of WP 6.5 “Legal and ethical acceptability of Symbiotic AI applications” of the spoke 6 “Symbiotic AI”, lead by the University of Bari.

## References

- [1] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019. URL: <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.1.html>.
- [2] A. Dyoub, S. Costantini, F. A. Lisi, Learning domain ethical principles from interactions with users, *Digital Society* 1 (2022). URL: <https://doi.org/10.1007/s44206-022-00026-y>. doi:10.1007/s44206-022-00026-y.
- [3] A. Dyoub, S. Costantini, F. A. Lisi, Towards ethical machines via logic programming, in: B. Bogaerts, E. Erdem, P. Fodor, A. Formisano, G. Ianni, D. Incezan, G. Vidal, A. Villanueva, M. D. Vos, F. Yang (Eds.), *Proceedings 35th International Conference on Logic Programming (Technical Communications), ICLP 2019 Technical Communications, Las Cruces, NM, USA, September 20-25, 2019, volume 306 of EPTCS, 2019*, pp. 333–339. URL: <https://doi.org/10.4204/EPTCS.306.39>. doi:10.4204/EPTCS.306.39.
- [4] A. Dyoub, S. Costantini, F. A. Lisi, Logic programming and machine ethics, in: F. Ricca, A. Russo, S. Greco, N. Leone, A. Artikis, G. Friedrich, P. Fodor, A. Kimmig, F. A. Lisi, M. Maratea, A. Mileo, F. Riguzzi (Eds.), *Proceedings 36th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2020, (Technical Communications) UNICAL, Rende (CS), Italy, 18-24th September 2020, volume 325 of EPTCS, 2020*, pp. 6–17. URL: <https://doi.org/10.4204/EPTCS.325.6>. doi:10.4204/EPTCS.325.6.
- [5] A. Dyoub, S. Costantini, I. Letteri, F. A. Lisi, A logic-based multi-agent system for ethical monitoring and evaluation of dialogues, in: A. Formisano, Y. A. Liu, B. Bogaerts, A. Brik, V. Dahl, C. Dodaro, P. Fodor, G. L. Pozzato, J. Vennekens, N. Zhou (Eds.), *Proceedings 37th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2021, Porto (virtual event), 20-27th September 2021, volume 345 of EPTCS, 2021*, pp. 182–188. URL: <https://doi.org/10.4204/EPTCS.345.32>. doi:10.4204/EPTCS.345.32.
- [6] M. West, R. Kraut, C. H. Ei, I’d blush if I could: closing gender divides in digital skills through education, Technical Report, EQUALS and UNESCO, 2019. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000367416>.
- [7] Division for Gender Equality, Artificial intelligence and gender equality: key findings of UNESCO’s Global Dialogue, Technical Report, UNESCO, 2020. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000374174>.
- [8] C. Collett, G. Neff, L. G. Gomes, The effects of AI on the working lives of women, Technical Report, UNESCO, Inter-American Development Bank, and Organisation for

Economic Co-operation and Development, 2022. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000380861>.

- [9] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. Langhans, M. Tegmark, F. F. Nerini, The role of artificial intelligence in achieving the sustainable development goals, *Nature Communications* 11 (2020). URL: <https://doi.org/10.1038/s41467-019-14108-y>. doi:10.1038/s41467-019-14108-y.
- [10] S. Badaloni, F. A. Lisi, Towards a gendered innovation in AI (short paper), in: G. Vizzari, M. Palmonari, A. Orlandini (Eds.), *Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020)*, Anywhere, November 27th, 2020, volume 2776 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 12–18. URL: <http://ceur-ws.org/Vol-2776/paper-2.pdf>.
- [11] D. J. Haraway, *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, Routledge/New York, 1991.
- [12] R. Braidotti, *The Posthuman*, Polity Press, Cambridge, 2013.
- [13] C. D’Ignazio, L. Klein, *Data Feminism*, The MIT Press, 2020.
- [14] V. Dignum, Relational artificial intelligence, *CoRR abs/2202.07446* (2022). URL: <https://arxiv.org/abs/2202.07446>. arXiv:2202.07446.
- [15] S. Badaloni, A. Rodà, Gender knowledge and artificial intelligence, in: G. Boella, F. A. D’Asaro, A. Dyoub, G. Primiero (Eds.), *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, University of Udine, Udine, Italy, 2022, CEUR-WS.org, 2023, pp. ?–?