

Responsible Research and Innovation for a Trustworthy AI: Recent activities in Ethics and Gender at Uniba

Francesca Alessandra Lisi

Dip. Informatica - Università degli Studi di Bari “Aldo Moro”

FrancescaAlessandra.Lisi@uniba.it

Abstract

This paper is a brief account of recent activities in Responsible Research and Innovation (RRI) that cover two thematic elements of RRI: Ethics and Gender. These themes are also central in the guidelines for a trustworthy AI.

1 Introduction

Responsible Research and Innovation (RRI) is an approach 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 has gained increasing EU policy relevance in the last decade, in particular within the European Commission’s Science in Society programme, in the context of the Horizon 2020 Strategy.¹ RRI involves 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 Open Access. RRI principles are being applied especially to disruptive technologies with the aim of revealing how these technologies can be socially and economically transformative, and to identify the potential downsides of the “disruption” it might bring about. This is definitely the case of Artificial Intelligence (AI).

The growing number of successful AI applications raises indeed several new issues, notably the need to increase the degree of trust in AI technologies. According to the Guidelines presented by the High-Level Expert Group on Artificial Intelligence of the European Commission [HLAI, 2019],² *trustworthy AI* should be: (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 covered in the Guidelines. 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. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

RRI and trustworthy AI have therefore an important intersection, notably the two themes of Ethics and Gender.

In this short paper we briefly describe some activities - relevant to the workshop “AI Responsabile ed Affidabile” - which have involved the participation of members of the CINI AIIS lab unit of Bari over the past two years.

2 Theme 1: Ethics

Ethics come into play in many AI applications. For instance, the problem of evaluating the ethical behaviour of AI-based chatbots in customer service has been addressed by [Dyoub *et al.*, 2019; Dyoub *et al.*, 2020; Dyoub *et al.*, 2021]. 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. 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.

Ethics has been the subject of a couple of initiatives organized by the COST Action “DIGital FOREnsics: evidence Analysis via intelligent Systems and Practices” (DigForASP).³ DigForASP explores the potential of the application of Artificial Intelligence and Automated Reasoning in the Digital Forensics field to create synergies between these fields. The Action is still running and will end later this year.

The former event, entitled “DigForEthics: Conversations on Digital Forensics, Ethics, Law and Artificial Intelligence”, was organized by Prof. Viviana MASCARDI from the University of Genova, Italy, and Prof. Juan Carlos NIEVES from Umeå University, Sweden.⁴ The event, held online on January 28-29, 2021, attracted more than 140 registered participants with a peak of 70 online attendees on the first day.

¹RRI - <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation>

²<https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.1.html>

³DigForASP - <https://digforasp.uca.es/>

⁴<https://www.cost.eu/when-artificial-intelligence-meets-digital-forensics-ethics-and-law/>

Overall, the level of discussion was very high and confirms that events like DigForEthics, where experts from different disciplines sit together and discuss not only general, but also very focussed and specific problems, are extremely needed. In particular, Prof. Francesca Alessandra LISI from the University of Bari, Italy, shared with the audience some recent results and activities in Trustworthy AI, with a focus on the gender dimension.

The latter event, entitled “Automated Decision Making, AI systems and individual rights”, was organized by Dr. Chiara Gallese Nobile from Eindhoven University of Technology, NL, Prof. David Billard from the University of Applied Sciences in Geneva, CH, Dr. Elena Falletti from Carlo Cattaneo University LIUC, IT, and Prof. Francesca A. Lisi from the University of Bari “Aldo Moro”, IT.⁵ The conference, held online on July 15-16, 2021, explored the ethical, legal, and technical issues of ADM systems, such as the need for transparent and explainable AI, the difficulties in eliminating biases in data sets, and the consequences of using ADM systems in the legal and medical field. The issue of trust in AI systems has also been extensively discussed.

3 Theme 2: Gender

With reference to gender, a number of initiatives have been recently undertaken, among which the ACM WoMENcourage 2019 workshop “Gendering ICT”⁶ addressed the twofold problem of including the gender dimension in computer science/engineering and increasing the presence of women in the field. The workshop also stressed the importance of paying more attention to how data are collected, processed and organized in machine learning applications. More focused on gender issues in AI, the online events “Questioni di genere in Intelligenza Artificiale”⁷ and “Intelligenza Artificiale e impatti di genere: una questione urgente”⁸ were targeted to the general public and to the management, respectively. Both were organized under the patronage of the Italian Association for Artificial Intelligence (AIXIA). The former took place on November 25, 2020, as part of the the annual conference of AIXIA, and was organized and chaired by Francesca A. Lisi (UNIBA). The latter was held on July 13, 2021, in collaboration with ESTE Edizioni and with the participation of Francesca A. Lisi as invited speaker.

In July 2021 the University of Bari (UNIBA) launched the initiative “Horizon Europe Seeds”, to support project proposals of highly interdisciplinary research that can be the ‘seed’ for prospective applications to the forthcoming calls of the Horizon Europe programme in one of the 6 clusters of Pillar II. In particular, “Cluster 2: Culture, Creativity and Inclusive society” aims to strengthen European democratic values, including rule of law and fundamental rights, safeguarding our

cultural heritage, and promoting socio-economic transformations that contribute to inclusion and growth.

Among the projects funded by Horizon Europe Seeds for Cluster 2, 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, institutional documents will be analyzed with Natural Language Processing 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 to derive policy rules. The team of AI researchers involved in the project are Francesca A. Lisi (Key Area person), Marco de Gemmis, and Pierpaolo Basile.

4 Challenges

As for the **Ethics** theme, the main problem is the current lack of benchmarks for evaluating and assessing the adherence of AI systems to the AI Ethics guidelines, as stressed also during the discussion at the AAAI Spring Symposium 2021 on implementing AI Ethics. Concerning the work on ethical chatbots in [Dyoub *et al.*, 2021], a web application⁹ has been implemented to collect data, regarding unethical scenarios. The aim is to produce a big dataset to train the proposed customer service chatbot. Also, the MAS prototype has a great potential for future implementations of ethical chatbots in different domains. Indeed, simulation by means of a MAS might help to get better insights into the dynamics of a corresponding real-world system, and to assess the practical challenges and limitations of building such a system.

A major challenge concerning the **Gender** theme 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.¹⁰ Considering these approaches may add valuable dimensions to research. In the case of AI, this means 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 both from a methodological point of view and from the applicative one. In [Badaloni e Lisi, 2020], it is 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<https://www.cost.eu/artificial-intelligence-systems-and-individual-rights-where-are-we-heading-to/>

⁶<http://www.di.uniba.it/~lisi/genderingICT/>

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

⁸<https://www.storicoeventi.este.it/eventi-per-data/1038-convegno-virtuale-intelligenza-artificiale-e-impatti-di-genere/-una-questione-urgente.html>

⁹<http://ethicalchatbot.sytes.net/eng/>

¹⁰<http://genderedinnovations.stanford.edu/index.html>

References

- [Badaloni e Lisi, 2020] Silvana Badaloni e Francesca A. Lisi. Towards a gendered innovation in AI (short paper). In Giuseppe Vizzari, Matteo Palmonari, e Andrea Orlandini, editors, *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*, pages 12–18. CEUR-WS.org, 2020.
- [Dyoub *et al.*, 2019] Abeer Dyoub, Stefania Costantini, e Francesca A. Lisi. Towards ethical machines via logic programming. In Bart Bogaerts, Esra Erdem, Paul Fodor, Andrea Formisano, Giovambattista Ianni, Daniela Incelezan, Germán Vidal, Alicia Villanueva, Marina De Vos, e Fangkai Yang, editors, *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*, pages 333–339, 2019.
- [Dyoub *et al.*, 2020] Abeer Dyoub, Stefania Costantini, e Francesca A. Lisi. Logic programming and machine ethics. In Francesco Ricca, Alessandra Russo, Sergio Greco, Nicola Leone, Alexander Artikis, Gerhard Friedrich, Paul Fodor, Angelika Kimmig, Francesca A. Lisi, Marco Maratea, Alessandra Mileo, e Fabrizio Riguzzi, editors, *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*, pages 6–17, 2020.
- [Dyoub *et al.*, 2021] Abeer Dyoub, Stefania Costantini, Ivan Letteri, e Francesca A. Lisi. A logic-based multi-agent system for ethical monitoring and evaluation of dialogues. In Andrea Formisano, Yanhong Annie Liu, Bart Bogaerts, Alex Brik, Verónica Dahl, Carmine Dodaro, Paul Fodor, Gian Luca Pozzato, Joost Vennekens, e Neng-Fa Zhou, editors, *Proceedings 37th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2021, Porto (virtual event), 20-27th September 2021*, volume 345 of *EPTCS*, pages 182–188, 2021.
- [HLAI, 2019] HLAI. Ethics guidelines for trustworthy AI. Technical report, European Commission, 2019.