

BEWARE-22: Bringing together researchers to address the logical, ethical, and epistemological challenges of AI

Guido Boella¹, Fabio Aurelio D'Asaro^{2,*}, Abeer Dyoub³ and Giuseppe Primiero⁴

¹*Department of Computer Science, University of Turin, Turin, Italy*

²*Ethos Group, Department of Human Sciences, University of Verona, Verona, Italy*

³*DISIM, University of L'Aquila, L'Aquila, Italy*

⁴*LUCI Group, Department of Philosophy, University of Milan, Milan, Italy*

Abstract

The BEWARE-22 workshop, held on December 2, 2022 in Udine, Italy, focused on emerging ethical aspects of artificial intelligence, with a particular emphasis on bias, risk, explainability, and the role of logic and logic programming. The invited speaker, Francesca Alessandra Lisi, gave a talk on “Ethics & Gender for a Responsible Research and Innovation in AI,” exploring the intersection of ethics and gender in the context of responsible research and innovation in artificial intelligence. The workshop program consisted of three sessions: “Logic for AI”, “Technical Approaches to XAI”, and “Conceptual Views,” which this short preface aims to describe. In total, 13 papers were accepted for the workshop, with 5 accepted as long papers and 8 as short papers. The proceedings include 12 papers out of the 13 from the workshop, plus an invited abstract, and will hopefully serve as a valuable resource for researchers and practitioners working on the ethical aspects of AI, inspiring further discussions and collaborations in this critical area of research.

Keywords

Ethical AI, Explainable AI, Logic, Logic Programming

1. Introduction

It is with great pleasure that we present the proceedings of the *BEWARE-22* workshop, held on December 2, 2022 in Udine, Italy, co-located with the *AIXIA 2022* conference. The *BEWARE-22* Workshop was a forum focused on discussing the ethical aspects of Artificial Intelligence (AI), with a particular emphasis on bias, risk, explainability, and the role of logic and logic programming (also see the website at <http://sites.google.com/view/beware2022>). It was the result of merging the *BRIO Workshop* (short for *Bias, Risk and Opacity in AI*, linked to 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


*Corresponding author.


✉ guido.boella@unito.it (G. Boella); fabioaurelio.dasaro@univr.it (F. A. D'Asaro); abeer.dyoub@univaq.it (A. Dyoub); giuseppe.primiero@unimi.it (G. Primiero)

🌐 <http://sites.google.com/view/fdasaro> (F. A. D'Asaro); <https://www.abeerdyoub.com> (A. Dyoub);

<https://sites.unimi.it/gprimiero/> (G. Primiero)

🆔 0000-0002-2958-3874 (F. A. D'Asaro); 0000-0003-0329-2419 (A. Dyoub); 0000-0003-3264-7100 (G. Primiero)

 © 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)

PRIN2020 (2020SSKZ7R) BRIO, see <https://sites.unimi.it/brio/> for the website), the *2nd Edition of the ME&E-LP Workshop* (short for *Machine Ethics & Explainability - the Role of Logic Programming*, see <http://sites.google.com/view/meande2021> for the website of the first edition and [1] for the joint proceedings volume of workshops at ICLP 2021), and the *AI AWARE Workshop* (short for *Ethics and AI, a two-way relationship*, linked to the AI AWARE Project, see <https://ai-aware.unito.it> for the website). BEWARE-22 aimed to bring together researchers from various disciplines, including AI, philosophy, ethics, epistemology, and social science, to promote collaborations and discussions on the development of trustworthy AI methods and solutions that are technologically reliable and socially acceptable. It addressed issues of logical, ethical, and epistemological nature in AI through the use of interdisciplinary approaches and invited submissions from computer scientists, philosophers, economists, and sociologists interested in discussing contributions related to the formulation of epistemic and normative principles for AI, their conceptual representation in formal models, and their development in formal design procedures and computational implementations.

2. The Invited Talk and Abstract

We were honored to host Francesca Alessandra Lisi from the University of Bari as our invited speaker, who gave an invited talk on “Ethics & Gender for a Responsible Research and Innovation in AI”. Francesca’s talk explored the intersection of ethics and gender in the context of responsible research and innovation in artificial intelligence. The topic of ethics and gender is of increasing importance in the field of AI, as the development and deployment of AI systems can have significant impacts on society and individuals. The talk provided valuable insights into the ways in which AI research and innovation can be guided by ethical considerations and a commitment to diversity and inclusivity. We are also pleased to include an extended abstract of Francesca’s talk in the proceedings of the workshop. The extended abstract [2], which we highly recommend checking out, discusses the role of ethics and gender in the development of artificial intelligence. It highlights the need for a responsible approach to AI, particularly in light of the potentially disruptive effects of technology on society. This approach, known as Responsible Research and Innovation (RRI), involves considering ethical and gender-related issues in the design and development of AI. The abstract also discusses the Ethics Guidelines for a Trustworthy AI developed by the European Commission. These guidelines state that AI should be lawful, ethical, and robust in order to be trustworthy. The abstract goes on to describe a variety of activities related to AI ethics and gender that have been carried out by the AI community, with a particular focus on initiatives promoted by the Italian Association for Artificial Intelligence (AIxIA). Francesca suggests that future research on AI ethics could be informed by contemporary feminist theories, which can provide valuable insights into the ways in which power dynamics, particularly those related to gender, can shape the development and use of technology. By considering these issues, researchers can work towards designing AI systems that are more inclusive and just. Francesca also suggests that efforts to engage with diverse stakeholders, including those from underrepresented groups, will be crucial in ensuring that the development of AI reflects the needs and values of society as a whole.

3. Contents of the Proceedings

We received a total of 13 submissions, all of which were accepted, 7 as long papers and 6 as short papers. However, one research group eventually opted out and their paper is not included in the workshop proceedings. Therefore, this volume only includes 12 papers (8 long and 4 short) and 1 invited abstract for the invited talk. The proceedings of BEWARE-22 mirror its program, which consisted of three main sessions: “Logic for AI,” “Technical Approaches to XAI,” and “Conceptual Views.” To guide the reader through this proceedings volume, we will now delve into each session and paper in greater detail.

3.1. “Logic for AI” Session

The “Logic for AI” session featured research on the use of logic and logical reasoning in the development and application of artificial intelligence systems. Topics included proof-checking bias in labeling methods, counting propositional logic, logics for binary-input classifiers and their explanations, and reasoning about algorithmic opacity. In particular, the paper [3] introduces a typed natural deduction system to formally verify the presence of bias in automatic labeling methods. The system interprets data as terms and labels as types, with contexts encoding probability distributions on training data. Bias is understood as the divergence of expected probabilistic labeling by a classifier trained on opaque data from the fairness constraints set by a transparent dataset; the paper [4] discusses the use of counting propositional logic in relation to randomized computation, and examines the expressive power of its univariate fragment. The paper also presents a method for measuring the probability of counting formulas and shows that the logic can be used to simulate certain events associated with dyadic distribution; the paper [5] presents work on modal logics for binary-input classifiers and their explanations. The logics are able to represent classifiers that propositional logic cannot, and they are applied to explainable artificial intelligence. Finally, in the paper [6], Ekaterina Kubyshkina and Mattia Petrolo provide an epistemological characterization of the opacity of algorithms based on a tripartite analysis of their components. They introduce a formal framework using the neighborhood semantics for evidence logic to reason about an agent’s epistemic attitudes toward an algorithm and investigate the conditions that must be met to achieve epistemic transparency.

3.2. “Technical Approaches to XAI” Session

The “Technical Approaches to XAI” session focused on technical challenges and solutions related to explainable AI and the development of trustworthy and transparent AI systems. Papers in this session explored issues such as bias and fairness in learning systems, using inductive logic programming to approximate neural networks for preference learning, and addressing the dataset shift problem in brain-computer interface applications. More specifically, the paper [7] proposes a framework for generating synthetic data with specific types of bias and their combinations. The authors discuss the relationship between biases and moral and justice frameworks, and use their synthetic data generator to perform experiments on different scenarios with various bias combinations to analyze the impact of biases on performance and fairness metrics in machine learning models; the paper [8] approaches the definitions and

analysis of fairness and bias in learning systems from a generative perspective, focusing on the role of data generators in the learning process. The paper discusses the challenges of defining and measuring fairness and bias in learning systems and proposes a framework for analyzing these concepts based on the generation process; the paper [9] explores the use of Inductive Logic Programming to explain black-box models, specifically neural networks, when they are used to learn user preferences. The authors create a dataset of user preferences, train a set of NNs on this data, and perform experiments to investigate how ILP can globally approximate these Neural Networks. They also experiment with using Principal Component Analysis to reduce the dimensionality of the dataset while maintaining transparency in the explanations; the paper [10] discusses the problem of dataset shift in the context of brain-computer interface systems, where the data used for training and testing can come from different distributions and result in poor generalization performance. The authors propose a framework to improve the robustness and reliability of BCI systems and mitigate the dataset shift problem; the paper [11] discusses fairness and bias in artificial intelligence and proposes a framework for investigating and mitigating bias in explainable AI systems. The authors discuss the role of data quality, transparency, and accountability in achieving fairness and describe a case study of bias in an explainable AI system for credit scoring.

3.3. “Conceptual Views” Session

The “Conceptual Views” session examined the broader ethical and philosophical implications of AI, including the use of enthymematic counterfactuals to explain predictions and the role of gender knowledge in AI. The session also included a survey of philosophical work on explanation in the context of explainable AI. In the paper [12], the authors argue that counterfactual explanations for high-stakes decisions informed by computer models should be based on domain-specific and commonsensical principles that can be negotiated. They present a method for incorporating these principles into an explanatory dialogue system using enthymematic reasoning; the paper [13] provides a roadmap of recent work on the concept of explanation in the field of explainable artificial intelligence from the perspective of philosophical ideas on explanations and models in science; finally, the paper [14] discusses gender-related biases in machine learning-based systems and presents the experience of the “Gender Knowledge and Ethics in Artificial Intelligence” course offered at the School of Engineering at the University of Padova.

4. Conclusion

Overall, the workshop was a huge success, with high-quality papers and massive participation from researchers and practitioners in the field. We would like to express our sincere gratitude to all of the participants who contributed to the workshop’s success, namely: Melissa Antonelli, Andrea Apicella, Silvana Badaloni, Alexander Berman, Jean-Philippe Bernardy, Alessandro Bogliolo, Ellen Breitholtz, Alessandro Castelnovo, Andrea Cosentini, Riccardo Crupi, Fabio Aurelio D’Asaro, Serge Dolgikh, Daniele Fossemò, Christine Howes, Nicole Inverardi, Francesco Isgrò, Aleks Knoks, Ekaterina Kubyshkina, Xinghan Liu, Emiliano Lorini, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, Filippo Mignosi, Mattia Petrolo, Roberto Prevete, Giuseppe

Primiero, Luca Raggioli, Thomas Raleigh, Daniele Regoli, Antonio Rodà, Matteo Spezialetti, Muhammad Suffian. We would also like to thank our PC Members who contributed to the success of our workshop with their timely and precious work, namely: Damiano Azzolini (Università degli Studi di Ferrara), Massimiliano Badino (Università degli Studi di Verona), Paolo Baldi (Università degli Studi di Milano), Guido Boella (Università di Torino), Daniele Chiffi (Politecnico di Milano), Stefania Costantini (Università degli Studi dell'Aquila), Marcello D'Agostino (Università degli Studi di Milano), Fabio Aurelio D'Asaro (Università degli Studi di Verona), Giovanni De Gasperis (Università degli Studi dell'Aquila), Luigi Di Caro (Università di Torino), Abeer Dyoub (Università degli Studi dell'Aquila), Rino Falcone (Institute of Cognitive Sciences and Technologies-CNR), Roberta Ferrario (ISTC-CNR), Mattia Fumagalli (Università di Bolzano), Ekaterina Kubyshkina (Università degli Studi di Milano), Francesca Alessandra Lisi (Università degli Studi di Bari Aldo Moro), Ludovica Marinucci (ISTC-CNR), Michela Milano (Università di Bologna), Francesco Pedrazzoli (Università degli Studi di Verona), Daniele Porello (Università degli Studi di Genova), Davide Posillipo (Alkemy), Francesca Pratesi (ISTI-CNR Pisa), Roberto Prevete (Università degli Studi di Napoli Federico II), Giuseppe Primiero (Università degli Studi di Milano), Giovanni Sartor (Università di Bologna), Teresa Scantamburlo (Università Ca' Foscari), Viola Schiaffonati (Politecnico di Milano), Matteo Spezialetti (Università degli Studi dell'Aquila), Guglielmo Tamburrini (Università degli Studi di Napoli Federico II) and Alberto Termine (Università degli Studi di Milano).

We hope that the proceedings of this workshop will serve as a valuable resource for researchers and practitioners working on the ethical aspects of AI, and that they will inspire further discussions and collaborations in this critical area of research.

References

- [1] J. Arias, F. A. D'Asaro, A. Dyoub, G. Gupta, M. Hecher, E. LeBlanc, R. Peñaloza, E. Salazar, A. Saptawijaya, F. Weitkämper, J. Zangari (Eds.), Proceedings of the International Conference on Logic Programming 2021 Workshops co-located with the 37th International Conference on Logic Programming (ICLP 2021), Porto, Portugal (virtual), September 20th-21st, 2021, volume 2970 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: <http://ceur-ws.org/Vol-2970>.
- [2] F. A. Lisi, Ethics and gender for responsible research and innovation in AI, in: [15], 2023.
- [3] G. Primiero, F. A. D'Asaro, Proof-checking bias in labeling methods, in: [15], 2023.
- [4] M. Antonelli, Two remarks on counting propositional logic, in: [15], 2023.
- [5] X. Liu, E. Lorini, Logics for binary-input classifiers and their explanations, in: [15], 2023.
- [6] M. Petrolo, E. Kubyshkina, Reasoning about algorithmic opacity, in: [15], 2023.
- [7] A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli, A. Cosentini, Investigating bias with a synthetic data generator: Empirical evidence and philosophical interpretation, in: [15], 2023.
- [8] S. Dolgikh, Fairness and bias in learning systems: a generative perspective, in: [15], 2023.
- [9] D. Fossemò, F. Mignosi, L. Raggioli, M. Spezialetti, F. D'Asaro, Using inductive logic programming to globally approximate neural networks for preference learning: challenges and preliminary results, in: [15], 2023.

- [10] A. Apicella, F. Isgrò, R. Prevete, XAI approach for addressing the dataset shift problem: BCI as a case study, in: [15], 2023.
- [11] M. Suffian, A. Bogliolo, Investigation and mitigation of bias in explainable AI, in: [15], 2023.
- [12] A. Berman, E. Breitholtz, C. Howes, J.-P. Bernardy, Explaining predictions with enthymematic counterfactuals information, in: [15], 2023.
- [13] A. Knoks, T. Raleigh, XAI and philosophical work on explanation: A survey, in: [15], 2023.
- [14] S. Badaloni, A. Rodà, Gender knowledge and artificial intelligence, in: [15], 2023.
- [15] G. Boella, F. A. D'Asaro, A. Dyoub, G. Primiero (Eds.), 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 Workshop Proceedings, 2023.