XAI.it 2024 - Preface to the Fifth Italian Workshop on eXplainable Artificial Intelligence

Marco Polignano[1], Cataldo Musto[1], Roberto Pellungrini[2], Erasmo Purificato[3,‡], Giovanni Semeraro[1] and Mattia Setzu[2]

[1]University of Bari Aldo Moro, Italy
[2]Scuola Normale Superiore, Italy
[3]Joint Research Centre, European Commission, Ispra, Italy

Abstract

As Artificial Intelligence (AI) systems become integral to daily life, ensuring transparency and interpretability in their decision-making processes is critical. The General Data Protection Regulation (GDPR) has underscored users' right to understand how AI-driven systems make decisions that affect them. However, the pursuit of model performance often compromises explainability, creating a tension between achieving high accuracy and maintaining transparency. Core research questions focus on reconciling the high performance of LLMs and other AI models with interpretability requirements. Emerging research focuses on designing transparent systems, understanding the effects of opaque models on users, developing explanation strategies, and enhancing user control over AI behaviors. The workshop on eXplainable AI (XAI.it) provides a platform for addressing these challenges, fostering collaboration within the XAI community to explore novel solutions and share insights across this evolving multifaceted field.

Keywords

eXplainable AI, Biases, Trustworthiness, Large Language Models, LLMs, XAI

1. Motivations and Scientific Relevance

We are experiencing a new "AI summer" as artificial intelligence algorithms become widely adopted across a diverse range of fields, from media and entertainment to healthcare, finance, and legal decision-making. While early AI systems were relatively straightforward and interpretable, the rise of complex models, particularly those based on Deep Neural Networks (DNNs), has led to powerful yet opaque methodologies. These models' effectiveness is offset by their complexity—characterized by deep layers and a vast number of parameters—which often makes them difficult to understand or scrutinize. As intelligent systems are increasingly used in sensitive domains, the adoption of black-box models without clear interpretability mechanisms is both impractical and potentially risky.

The advent of Large Language Models (LLMs) has further amplified the challenges of transparency and interpretability. LLMs are highly effective at language-related tasks, such as text generation, summarization, and language translation, yet their internal decision-making processes are often difficult to interpret due to the scale and complexity of their architectures. The opaque nature of LLMs can limit user trust and raise concerns about the potential for bias, misinformation, and unintended consequences

XAI.it - 5th Italian Workshop on eXplainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024

‡The author contributed to this work while affiliated with Otto von Guericke University Magdeburg, Germany. The view expressed in this paper is purely that of the author and may not, under any circumstances, be regarded as an official position of the European Commission.

†All authors contributed equally.

marco.polignano@uniba.it (M. Polignano); cataldo.musto@uniba.it (C. Musto); roberto.pellungrini@sns.it (R. Pellungrini); erasmo.purificato@acm.org (E. Purificato); giovanni.semeraro@uniba.it (G. Semeraro); mattia.setzu@unipi.it (M. Setzu)
https://marcopoli.github.io/ (M. Polignano); https://swap.di.uniba.it/members/musto.cataldo/ (C. Musto); https://kdd.isti.cnr.it/people/pellungrini-roberto (R. Pellungrini); https://erasmopurif.com/ (E. Purificato); https://swap.di.uniba.it/members/semeraro.giovanni/ (G. Semeraro); hhttps://kdd.isti.cnr.it/people/setzu-mattia (M. Setzu)
0000-0002-3939-0136 (M. Polignano); 0000-0001-6089-928X (C. Musto); 0000-0002-1366-9833 (R. Pellungrini); 0000-0002-5506-3020 (E. Purificato); 0000-0001-6883-1853 (G. Semeraro); 0000-0001-8351-9999 (M. Setzu)

in high-stakes applications. As LLMs increasingly impact areas such as education, healthcare, and content moderation, the need for models that are not only accurate but also explainable becomes paramount. Developing interpretability techniques for LLMs that enable users to understand and control these models' outputs is essential for ensuring responsible AI deployment. Conventional metrics for evaluating AI performance often prioritize accuracy, inadvertently favoring these opaque models at the expense of transparency. This trade-off has been highlighted by recent regulations and initiatives, such as the General Data Protection Regulation (GDPR) and DARPA's eXplainable AI Project, which underscore a growing need for AI methodologies that are both effective and interpretable. These initiatives have reinforced the user's right to transparency, emphasizing that AI-driven decisions must be comprehensible to build trust and ensure accountability.

The motivation behind this workshop is to tackle the crucial question: how can we reconcile the effectiveness of advanced AI systems with the imperative for transparency and interpretability? This question opens up several important research avenues, including developing transparent and interpretable models that retain high performance, enabling humans to better understand and trust AI-based methods, and establishing ways to assess the transparency and explainability of AI systems as a whole. The workshop provides a platform for the Italian research community to engage in these critical discussions, share innovative approaches, and address the pressing challenges in explainable AI and LLM transparency.

2. Accepted Papers

We believe the program offers a well-balanced exploration of the diverse topics within the field of eXplainable AI. This year's program is further enhanced by a keynote presentation focused on the evolving role of XAI in the era of Large Language Models (LLMs). The accepted papers cover a wide array of contributions, from proposing novel methodologies to enhance the interpretability of AI systems to developing new applications that embody eXplainable AI principles. A total of 8 submissions were received for XAI.it 2024, with 6 selected for inclusion in the proceedings, plus a position paper by organizers:

- *Daehyun Yoo* and *Caterina Giannetti*. Ethical AI Systems and Shared Accountability: The Role of Economic Incentives in Fairness and Explainability [1].
- *Silvia D'Amicantonio*, *Mishal Kizhakkam Kulangara*, *Het Darshan Mehta*, *Shalini Pal*, *Marco Levantesi*, *Marco Polignano*, *Erasmo Purificato*, and *Ernesto William De Luca*. A Comprehensive Strategy to Bias and Mitigation in Human Resource Decision Systems [2].
- *Leonardo Dal Ronco* and *Erasmo Purificato*. ExplainBattery: Enhancing Battery Capacity Estimation with an Efficient LSTM Model and Explainability Features [3].
- *Ejdis Gjinika*, *Nicola Arici*, *Luca Putelli*, *Alfonso Emilio Gerevini*, and *Ivan Serina*. An Analysis on How Pre-Trained Language Models Learn Different Aspects [4].
- *Giovanna Castellano*, *Maria Grazia Miccoli*, *Raffele Scaringi*, *Gennaro Vessio*, and *Gianluca Zaza*. Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps [5].
- *Zhuofan Zhang* and *Herbert Wiklicky*. Probabilistic Abstract Interpretation on Neural Networks via Grids Approximation [6].
- (Position Paper) *Marco Polignano*, *Cataldo Musto*, *Roberto Pellungrini*, *Erasmo Purificato*, *Giovanni Semeraro*, and *Mattia Setzu*. XAI.it 2024: An Overview on the Future of eXplainable AI in the era of Large Language Models [7].

3. Program Committee

As a final remark, the program co-chairs would like to thank all the members of the Program Committee (listed below), as well as the organizers of the AIxIA 2024 Conference [1].

[1] https://aixia2024.events.unibz.it/

- *Roberto Confalonieri*, University of Padua
- *Ruggero G. Pensa*, University of Torino
- *Antonio Rago*, Imperial College London
- *Giuseppe Sansonetti*, Roma Tre University
- *Valerio Basile*, University of Turin
- *Claudio Pomo*, Politecnico di Bari
- *Roberto Capobianco*, Sapienza University of Rome
- *Salvatore Ruggieri*, Università di Pisa
- *Ludovico Boratto*, University of Cagliari
- *Mirko Marras*, University of Cagliari

Acknowledgments

This research is partially funded by PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] D.-H. Yoo, C. Giannetti, Ethical AI Systems and Shared Accountability: The Role of Economic Incentives in Fairness and Explainability, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[2] S. D'Amicantonio, M. K. Kulangara, H. D. Mehta, S. Pal, M. Levantesi, M. Polignano, E. Purificato, E. W. De Luca, A Comprehensive Strategy to Bias and Mitigation in Human Resource Decision Systems, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[3] L. Dal Ronco, E. Purificato, ExplainBattery: Enhancing Battery Capacity Estimation with an Efficient LSTM Model and Explainability Features, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[4] E. Gjinika, N. Arici, L. Putelli, A. E. Gerevini, I. Serina, An Analysis on How Pre-Trained Language Models Learn Different Aspects, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[5] G. Castellano, M. G. Miccoli, R. Scaringi, G. Vessio, G. Zaza, Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[6] Z. Zhang, H. Wiklicky, Probabilistic Abstract Interpretation on Neural Networks via Grids Approximatio, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.

[7] M. Polignano, C. Musto, R. Pellungrini, E. Purificato, G. Semeraro, M. Setzu, XAI.it 2024: An Overview on the Future of Explainable AI in the era of Large Language Models, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.