Workshop on Human-Interpretable AI

Gabriele Ciravegna
gabriele.ciravegna@polito.it
Dipartimento di Automatica e Informatica, Politecnico di Torino
Torino, Italy

Mateo Espinoza Zarlenga
University of Cambridge
Cambridge, UK

Pietro Barbiero
Università della Svizzera Italiana
Lugano, Switzerland

Francesco Giannini
Scuola Normale Superiore
Pisa, Italy

Zoreh Shams
University of Cambridge
Cambridge, UK

Damien Garreau
Julius-Maximilians-Universität Würzburg
Würzburg, Germany

Mateja Jamnik
University of Cambridge
Cambridge, UK

Tania Cerquitelli
Dipartimento di Automatica e Informatica, Politecnico di Torino
Torino, Italy

Abstract

This workshop aims to spearhead research on Human-Interpretable Artificial Intelligence (HI-AI) by providing: (i) a general overview of the key aspects of HI-AI, in order to equip all researchers with the necessary background and set of definitions; (ii) novel and interesting ideas coming from both invited talks and top paper contributions; (iii) the chance to engage in dialogue with prominent scientists during poster presentations and coffee breaks. The workshop welcomes contributions covering novel interpretable-by-design or post-hoc approaches, as well as theoretical analysis of existing works. Additionally, we accept visionary contributions speculating on the future potential of this field. Finally, we welcome contributions from related fields such as Ethical AI, Knowledge-driven Machine learning, Human-machine Interaction, applications in Medicine and Industry, and analyses from Regulatory experts.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Human-Interpretable AI, Interpretability, Explainability, HI-AI, XAI

ACM Reference Format:
Gabriele Ciravegna, Mateo Espinoza Zarlenga, Pietro Barbiero, Francesco Giannini, Zoreh Shams, Damien Garreau, Mateja Jamnik, and Tania Cerquitelli. 2024. Workshop on Human-Interpretable AI. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3637528.3671499

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671499

1 Introduction

Human-interpretable AI models [1] are playing an increasingly important role in Artificial Intelligence (AI). Today, a large part of the technologies employed by AI and SIGKDD researchers is based on Deep Neural Networks (DNNs). Yet, the lack of transparency of DNNs prevents a safe deployment of these models in critical contexts that significantly affect users. Consequently, decision-making systems based on deep learning are facing constraints and limitations from regulatory institutions [2], which increasingly demand transparency in AI models [3]. Even though standard eXplainable AI (XAI) emerged to address the need to interpret DNNs, several works are arguing that it may not have achieved its goal [4, 5].

To really explain DNN decision-making process, there is a growing consensus that human-interpretable explanations are required. Human-Interpretable AI (HI-AI) methods either provide post-hoc explanations by extracting the symbols that have been automatically learnt by the models (e.g., T-CAV [6]), or directly design intrinsically interpretable architectures (e.g., CBM [7]). Among other qualities, these explanations resemble better the way humans reason and explain [8], help to detect model biases [9], are more stable to perturbations [10], and can create more robust models [11].

2 Workshop Topics

Topics of interest include, but are not limited to, the following:

- **Explainable-by-design models**, novel approaches to creating machine learning and deep learning models that are intrinsically explainable or interpretable.
- **Post-hoc methods for Interpretable AI**, novel approaches on post-hoc interpretable AI. These include but are not limited to approaches working on higher-level features such as concepts.
- **Theoretical analyses** of existing methods, showing what existing interpretable methods can achieve both from an explanation and a generalization point of view.
- **Knowledge integration & Reasoning** methods injecting domain knowledge and reasoning methods into deep learning models to enhance their interpretability and performance.

- **AI Ethics** papers analysing implications of interpretable AI methods, discussing topics such as fairness, accountability, transparency, and bias mitigation in AI systems.
- **Human-machine Interaction** studies on innovative human-machine interaction systems, successfully exploiting interpretable AI models in their capability to provide both standard and counter-factual explanations.
- **Vision papers on XAI** discussing the possible evolutions of the XAI field or speculating potential interpretable system and applications with their implications.
- **Applications in Medicine and Healthcare** applications of interpretable AI methods in medical diagnosis, treatment planning, and healthcare decision-making.
- **AI in Industry** practical applications of interpretable AI methods in various safety-critical industrial sectors, such as transportation, finance and retail.
- **Legal and Regulatory dissertations** discussing and providing analysis of the legal challenges associated with interpretable AI, including compliance with data protection laws for transparent and accountable AI systems.

3 Program

This workshop aims to advance the research on HI-AI by offering a diverse program designed to enhance participants' knowledge, and foster collaboration and innovation. The following list contains the invited speakers who will give keynote talks at the HI-AI workshop, and the expected topics that their talks will cover. All invited speakers have already confirmed their presence.

- **Abbas Rahimi**, Research Staff Member at IBM Research Europe - Neuro-symbolic AI, Concept Embeddings.
- **Andrea Passerini**, Associate Professor at University of Trento - Concepts in AI and Interactive Machine Learning.
- **Sonali Parbhoo**, Assistant Professor at Imperial College London - Concept and causality.

Program Outline. Table 1 reports the workshop program. Firstly, we will give an overview of the key aspects of HI-AI to ensure all attendees have a solid understanding of the background concepts and terminology. Secondly, the workshop features three invited talks from experts in the field, who will share their insights and latest research findings. These talks will provide valuable perspectives and inspire new ideas. Thirdly, we will offer participants the chance to engage in dialogue with prominent scientists during a long coffee break with poster presentations, encouraging collaborations and knowledge-sharing. Also, the workshop program includes three contributed talks from selected contributions. We will recognize the most interesting contribution with a Best Workshop Paper Award. We have allocated 40 minutes for each invited talk, allowing for a 30-minute presentation followed by a 10-minute Q&A session. We allotted the same time for the poster sessions.

4 Paper Management

Paper management. We published the Call For Papers (CFP) on the workshop website[1]. The CFP focuses on short papers, which can be research papers, theoretical analysis papers, or vision papers.

[1]https://human-interpretable-ai.github.io/

8:50 – 9:00	Opening remarks
9:00 – 9:40	**Keynote**: Andrea Passerini
9:40 – 10:00	5 mins lightning talks (3 selected papers)
10:00 – 10:40	**Keynote**: Abbas Rahimi
10:40 – 11:30	Coffee & Posters
11:30 – 12:10	**Keynote**: Sonali Parbhoo
12:10 – 12:20	**Awards and Closing Remarks**

Table 1: Draft of the program outline.

In the case of research contributions, we asked paper authors to make their code and data openly available to ensure reproducibility. The review process has been double-blind. We have used OpenReview to ensure the final decisions for each paper are made by the organisers with no conflict of interest. All accepted papers will be published on the workshop website, which will remain active and accessible after the conference concludes. Additionally, we took contact with an external editor (CEUR-WS) to create an archival version of these papers for authors who wish to participate in a subsequent publication.

5 Program Commitee

We are very grateful to each of our program committee members for their hard reviewing work, namely Romain Giot, Eliana Pastor, Roberto Pellungrini, Eleonora Poeta, Gianluigi Lopardo, and Gizem Gezici, besides workshop chairs.

References

[1] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.
[2] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.
[3] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.
[4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.
[5] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.
[6] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.
[7] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.
[8] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. " help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.
[9] Rishabh Jain, Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Davide Buffelli, and Pietro Lio. Extending logic explained networks to text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8838–8857. Association for Computational Linguistics, 2022.
[10] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.
[11] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic explained networks. *Artificial Intelligence*, 314:103822, 2023.