

Preface to the 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight (XAI.it 2026)

Marco Polignano^{1,*}, Cataldo Musto¹, Ernesto William De Luca^{2,3}, Erasmo Purificato^{4,†} and Giovanni Semeraro¹

¹Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

²Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany

³Leibniz Institute for Educational Media | Georg Eckert Institute, Freisestr. 1, 38118 Braunschweig, Germany

⁴European Centre for Algorithmic Transparency (ECAT), Joint Research Centre, European Commission

Abstract

This volume collects the papers presented at XAI.it 2026, the 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight, held on October 6, 2026 in Perugia, Italy, as part of the 24th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2026). The workshop addresses how transparency, interpretability, and human oversight can be realized for large language models, foundation models, and agentic systems, and how they can be evaluated rigorously. This preface summarizes the motivation and scope of the workshop, the review process, the accepted contributions, and the program.

Keywords

Explainable AI, Generative AI, Agentic AI, Trustworthy AI, Human oversight

1. Motivation and Scope

Explainability has become a central requirement for the deployment of modern AI systems in safety-critical, socially sensitive, and decision-support settings. The shift from task-specific predictive models to large language models, foundation models, and increasingly autonomous agents sharpens this requirement, because the opacity of such systems makes it difficult for users, stakeholders, and regulators to understand how an output was produced, to contest it, or to assess the harms and biases it may carry. This opacity undermines trust, slows adoption, and conflicts with emerging governance frameworks that call explicitly for transparency, accountability, and human oversight.

XAI.it is a workshop series on explainable artificial intelligence organized in conjunction with the AIxIA conference. The 2026 edition, the sixth, extends the scope of the series to generative and agentic systems. It is organized around three research questions:

1. How can more transparent and more explainable agents, large language models, and machine learning models be designed while maintaining high performance?
2. How can humans be enabled to understand and appropriately trust AI-based systems, methods, and digital agents and assistants?
3. How can the overall transparency, explainability, ethical quality, and social impact of AI systems and models be evaluated?

XAI.it 2026: 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight, co-located with AIxIA 2026, October 6, 2026, Perugia, Italy

*Corresponding author.

†The views expressed in this paper are purely those of the author and may not, under any circumstances, be regarded as an official position of the European Commission.

✉ marco.polignano@uniba.it (M. Polignano); cataldo.musto@uniba.it (C. Musto); ernesto.deluca@ovgu.de (E. W. D. Luca); erasmo.purificato@acm.org (E. Purificato); giovanni.semeraro@uniba.it (G. Semeraro)

🌐 <https://swap.di.uniba.it/members/polignano.marco/> (M. Polignano); <https://swap.di.uniba.it/members/musto.cataldo/> (C. Musto); <https://www.hcai.ovgu.de/Staff/Head/Ernesto+William+De+Luca.html> (E. W. D. Luca); <https://erasmopurif.com/> (E. Purificato); <https://swap.di.uniba.it/members/semeraro.giovanni/> (G. Semeraro)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Current evaluation practice in explainable AI is fragmented. It is often limited to technical proxies such as fidelity and sparsity, and it rarely incorporates human-centred measures such as comprehension, trust calibration, contestability, or long-term societal effects. The workshop therefore also promotes multi-dimensional evaluation frameworks, benchmarks, and protocols that combine quantitative and qualitative methods and that integrate ethical and legal perspectives, including fairness, accountability, privacy, safety, and environmental impact.

The call for papers invited contributions on, among others, explainability for generative and agentic AI, trustable and transparent models, justification models, interpretable machine learning, strategies for explaining black-box decision systems, new explanation styles, evaluation of transparency and interpretability, theoretical and technical aspects of explanation, ethics, argumentation theory, natural language processing and human-machine interaction for explainable AI, fairness and bias auditing, privacy-preserving explanations, monitoring of system behaviour, and successful applications and demonstrators of interpretable AI.

2. Review Process

Authors could submit regular papers (7 to 12 pages plus references) or short and position papers (4 to 6 pages plus references) in the single-column CEUR-WS format, with a mandatory declaration on the use of generative AI. Submissions were made through EasyChair and were reviewed single-blind, that is, author names were included in the submissions. Each submission was evaluated by at least two members of the Program Committee on the basis of originality, significance, relevance, and technical quality. The submission deadline was extended from July 20 to July 27, 2026, authors were notified on August 24, 2026, and camera-ready versions were due on September 14, 2026.

The workshop received 7 submissions, all of which were accepted for presentation and publication in this volume after peer review: five as regular papers and two as short papers. All accepted papers are published in the AIXIA series of CEUR-WS.

3. Accepted Contributions

The accepted papers span three related themes.

Explaining and auditing generative models. Roy et al. present a retrieval-augmented generation pipeline that detects, explains, and rewrites multilingual dark-pattern text in a single structured inference step, returning the label, the psychological mechanism, the supporting evidence span, the targeted harm, and a consumer-safe rewrite. Martí-Pérez et al. study whether large language models can translate labelled time-series prototypes into concise natural-language rules, and evaluate the rules through forward simulation and a small human study, reporting dataset-dependent benefits and a risk of confidence gains without accuracy gains. Chakraborty and Rodà use contrastive activation steering to analyse, layer by layer and across English and Italian, the geometry of bias and refusal directions in Llama-2-7B-Chat and GPT-OSS-20B, showing that visible refusal can coexist with accessible stereotype-associated representations.

Auditability and assurance of perception systems. Caravani et al. introduce an OCR-guided occlusion sensitivity protocol that uses detected text segments as interpretable perturbation units for black-box auditing of document classifiers, and show that predictions of the evaluated models are partly driven by corpus-specific provenance cues. Rehman and De Gasperis propose TRACE-Pave, a framework for road crack assessment in which learned computation stops at perception and every quantity an engineer must defend is derived by recorded deterministic rules, together with an audit of the data pipeline and cases of confident errors accompanied by convincing saliency maps. Buhrmester et al. take the position that post-hoc explanations should not be treated as evidence of trustworthy behaviour unless embedded in an assurance process, and propose a conceptual reference framework

that connects explanations with uncertainty estimation, quality tests, plausibility checks, and human oversight in safety-critical automatic target recognition.

Human oversight and delegation. Chiatti et al. report an online study with 56 AI-literate participants and 392 video question answering trials, showing that post-interaction trust strongly predicts delegation choices while remaining distinct from them, and propose enacted delegation traces as a methodological framework for studying multimodal and agentic AI use.

4. Program

The workshop took place on October 6, 2026, from 10:30 to 12:30, in room T1 of the Accademia di Belle Arti in Perugia. The session comprised an opening, seven presentations of 15 minutes each (12 minutes of presentation and 3 minutes of questions), grouped by thematic affinity, and closing remarks. Updates to the schedule and rooms are published on the AIxIA 2026 workshops program page.

5. Post-proceedings

A selection of the best papers presented at the workshop could be invited to submit an extended version for publication in *Intelligenza Artificiale*, the International Journal of the Italian Association for Artificial Intelligence, published by IOS Press.

6. Acknowledgments

We thank all authors for their contributions and the members of the Program Committee for their careful and timely reviews: Alejandro Bellogín (Universidad Autónoma de Madrid), Ludovico Boratto (University of Cagliari), Allegra De Filippo (University of Bologna), Alessandro Giuliani (University of Cagliari), Michal Kompan (Kempelen Institute of Intelligent Technologies), Alberto Carlo Maria Mancino (Politecnico di Bari), Mirko Marras (University of Cagliari), Het Darshan Mehta (Otto von Guericke University Magdeburg), Stefania Montani (University of Piemonte Orientale), Antonio Rago (Imperial College London), Giuseppe Sansonetti (Roma Tre University), Irene Siragusa (University of Palermo), and Giuseppe Spillo (University of Bari Aldo Moro). We also thank the organizers of AIxIA 2026 and the University of Perugia for hosting the workshop.

Declaration on Generative AI

During the preparation of this preface, the editors used Claude (Anthropic) to grammar checks and sentences rewriting. The editors reviewed and edited the content and take full responsibility.