Advancing Trustworthy in AI: Mission and Research Lines at the TRAIL Lab

Stefano Buzi[1], Simona Cacace[1], Andrea Loreggia[2], Nadia Maccabiani[3], Giorgio Pedrazzi[1], Mattia Savardi[4], Alberto Signoroni[4] and Laura Zoboli[3]

[1]*Department of Law, University of Brescia*
[2]*Department of Information Engineering, University of Brescia*
[3]*Department of Economics and Management, University of Brescia*
[4]*Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia*

Abstract

The Trustworthy AI Lab (TRAIL) at the University of Brescia promotes the development of reliable, transparent, and ethically aligned artificial intelligence through a robust interdisciplinary approach. Grounded in international standards and policy frameworks—including those of the EU High-Level Expert Group on AI, the EU AI Act, UNESCO, and the OECD—TRAIL implements Z-Inspection®, a comprehensive, lifecycle-based methodology for evaluating the trustworthiness of AI systems. This paper presents TRAIL's mission, interdisciplinary structure, and five flagship initiatives: a Z-Inspection® pilot on the use of generative AI in higher education; a Z-Inspection® best-practice assessment related to COVID-19 case study on a lung severity prediction system; VIPPSTAR, an EU-funded project supporting visually impaired youth; DORIAN GRAY, an EU-funded project about digital medicine and healty ageing; and AI4Gov-X, an EU co-founded initiative to enhance the integration of AI in public governance. TRAIL's diverse team integrates legal, ethical, managerial and technical expertise to deliver validated algorithms, educational materials, policy frameworks, and regulatory tools that promote AI systems designed to be trustworthy by default.

Keywords

Artificial intelligence, Trustworthy AI, Interdisciplinarity

1. Introduction

Recent regulatory developments—most notably the EU AI Act—highlight the growing imperative for ex-ante risk management, ethical compliance, and transparency in the design and deployment of AI systems. The Trustworthy AI Lab (TRAIL) was established on the belief that fostering trust in AI requires sustained collaboration across disciplines, bringing together bioethicists, legal scholars, medical professionals, technologists, and engineers. Integrating this diverse expertise from the outset enables the creation of AI systems that are not only reliable and transparent, but also lawful and aligned with EU fundamental values. TRAIL's mission encompasses the development of risk-aware, evidence-based AI solutions, the promotion of inclusive stakeholder engagement, and the contribution to policy frameworks that support the safe and responsible use of AI.

To highlight the interdisciplinary scope and practical impact of TRAIL's work, Table 1 presents a comparative overview of its current flagship initiatives.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
stefano.buzi@unibs.it (S. Buzi); simona.cacace@unibs.it (S. Cacace); andrea.loreggia@unibs.it (A. Loreggia); nadia.maccabiani@unibs.it (N. Maccabiani); giorgio.pedrazzi@unibs.it (G. Pedrazzi); mattia.savardi@unibs.it (M. Savardi); alberto.signoroni@unibs.it (A. Signoroni); laura.zoboli@unibs.it (L. Zoboli)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-07

Table 1
Comparative overview of TRAIL's flagship initiatives

Project	Domain	Key Technologies	Outcomes
Generative AI in Higher Education	Education	Large Language Models	Best-practice guidelines; white paper; policy recommendations
COVID-19 BS-Net	Medical Diagnostics	Deep Learning; Post-hoc assessment	Validated model; ethical/legal audit report
VIPPSTAR	Pediatric Digital Health	Serious Gaming; Wearable Sensors; GDPR-compliant AI; Regulatory Sandbox	Holistic support framework; Ethical and regulatory insights and compliance of AI
DORIAN GRAY	Cardiology and Neurology	GDPR-compliant AI; Trustworthy AI	Holistic support framework; Ethical and regulatory compliance of AI
AI4Gov-X	Public Sector	AI frameworks for public sector; Trustworthy AI	Educational programs

2. Z-Inspection® Methodology

Z-Inspection®[1] is the core auditing framework employed by TRAIL[2] to assess the trustworthiness of AI systems across their entire lifecycle—from initial conception and design to deployment and continuous monitoring. Developed as a dynamic and holistic methodology, it has been formally described in [1] and is recognized in the OECD AI Policy Observatory[3].

The methodology is structured around three main phases shaped according to the Deming cycle: 1) the Set Up phase; 2) the Assess phase; and 3) the Resolve phase. These enable a comprehensive evaluation, both holistic and analytic, which addresses both technical performance and broader societal implications. Z-Inspection® builds upon the European Commission's High-Level Expert Group (HLEG) framework for trustworthy AI, which outlines seven key requirements: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability. In addition to these, Z-Inspection® extends the evaluative scope to include critical concerns such as democratic values, non-discrimination, market competition, and the risks associated with power concentration.

A defining feature of the methodology is its use of iterative socio-technical scenario analysis, in which multidisciplinary panels collaboratively examine context-specific use cases. This process facilitates the identification and mitigation of ethical, legal, and technical vulnerabilities at each stage of system development, ensuring a rigorous, evidence-informed pathway to responsible AI innovation.

[1] https://z-inspection.org/ - Last visited 30 May 2025
[2] http://trail.unibs.it - Last visited 30 May 2025
[3] https://oecd.ai/en/catalogue/tools/z-inspection - Last visited 30 May 2025

3. Key Research Initiatives

This section presents TRAIL's key research initiatives, illustrating how its multidisciplinary and interdisciplinary approach and methodological framework are applied across diverse real-world domains to promote trustworthy AI.

3.1. Generative AI in Higher Education

In collaboration with over 170 global stakeholders, TRAIL is applying the Z-Inspection® methodology to evaluate large language model (LLM)-based tools in the context of UNESCO's guidance on AI in education. The initiative focuses on case studies such as AI-assisted grading and curriculum design, examining critical dimensions including data bias, transparency, privacy, and the pedagogical implications of automated decision-making. The project aims to produce a set of tangible outputs, including best-practice guidelines for educators, a comprehensive white paper, peer-reviewed publications, and policy recommendations designed to inform international regulatory and educational frameworks for the responsible use of AI in learning environments.

3.2. COVID-19 Case Study: BS-Net for deployable pneumonia severity assessment and radiology resident training

In response to the COVID-19 emergency, we partnered with ASST Spedali Civili of Brescia to evaluate BS-Net, a deep neural network developed for lung severity scoring from chest X-ray images [2]. Deployed in December 2020 to support clinical triage, BS-Net played a key role in assisting radiologists and healthcare professionals in the early identification of critical cases.

Following its deployment, BS-Net[4] underwent a comprehensive post-hoc assessment using the Z-Inspection® methodology[5], aimed at evaluating its trustworthiness across multiple dimensions. The evaluation covered technical aspects such as model design and data integrity, as well as ethical and legal considerations, including informed patient consent and liability in triage decisions. This interdisciplinary audit, conducted under real-world, crisis conditions, underscored the necessity of continuous and collaborative evaluation processes for AI systems operating in high-stakes environments [3].

Building on the outcomes of this assessment, current research activities focus on the use of BS-Net as a training tool for radiology residents. In this context, particular attention is being given to the trustworthy use of AI in medical education, with an emphasis on transparency, explainability, and ethical alignment in clinical learning environments [4].

3.3. VIPPSTAR: AI for Visual Impairment

VIPPSTAR (Visually Impaired children and adolescents: bridging the gap with Personalized Prevention Strategies, Tools, Approaches, and Resources)[6] is a Horizon Europe project coordinated by the University of Brescia, involving 19 partners across 11 countries. The project aims to develop personalized health-improving strategies and digital tools tailored to the needs of visually impaired youth, addressing a critical gap in pediatric healthcare.

[4] https://brixia.github.io/ - Last visited 30 May 2025
[5] https://z-inspection.org/best-practices/ - Last visited 30 May 2025
[6] https://vippstar.eu/ - Last visited 30 May 2025

Within VIPPSTAR, TRAIL leads the creation of a regulatory sandbox aligned with the EU AI Act, specifically designed for pediatric digital health applications. This sandbox serves as a controlled environment for the rigorous testing, evaluation, and validation of AI-driven healthcare technologies, ensuring compliance with emerging legal and ethical standards. The initiative highlights TRAIL's commitment to responsible innovation in ethically sensitive and high-impact domains, fostering AI solutions that prioritize safety, transparency, and inclusivity for vulnerable populations.

3.4. DORIAN GRAY: AI and digital health in neurology and cardiology for healthy ageing

DORIAN GRAY (Bridging Cognitive Decline and Cardiovascular Health for Healthy Aging)[7] is a Horizon Europe project coordinated by the University of Brescia, involving 25 partners from 12 countries. This multidisciplinary consortium is dedicated to transforming the prevention and management of mild cognitive impairment (MCI) in patients with cardiovascular diseases (CVD), addressing a critical intersection between neurological and cardiovascular health.

Within this initiative, TRAIL plays a pivotal role in analyzing the regulatory landscape that governs the research and development of medical devices employed in the project. Moreover, TRAIL is responsible for evaluating the ethical, legal, and social implications arising from the integration of these technologies within the physician-patient relationship. By doing so, TRAIL ensures that the deployment of innovative health technologies aligns with regulatory requirements and respects the complexities of clinical practice and patient care.

3.5. AI4Gov-X: Shaping AI for Public Governance

We are involved in the AI4Gov-X project, a four-year initiative co-funded by the European Union's Digital Europe Programme and led by leading Higher Education Institutions. Officially launched at Politecnico di Milano in February 2025, AI4Gov-X aims to enhance the integration of Artificial Intelligence in public governance while upholding democratic values.

TRAIL supports the design of educational modules within the AI4Gov-X Master's program, focusing on embedding trustworthy AI principles into the curriculum for future public sector leaders. Through these efforts, TRAIL plays a pivotal role in fostering responsible AI adoption in public services across Europe.

4. Conclusion

The TRAIL Lab, an interdepartmental research center at the University of Brescia, is at the forefront of advancing trustworthy AI through multidisciplinary and interdisciplinary approaches. By synthesizing legal, ethical, and technical perspectives, TRAIL delivers a human-centric, evidence-based framework that translates fundamental principles into practical applications. Its work spans critical sectors—from healthcare and assistive technologies to the deployment of generative AI in education—illustrating how rigorous assessment can drive responsible innovation. TRAIL collaborates closely with regulators and international bodies, contributing to the design of policy frameworks and regulatory sandboxes that ensure AI technologies remain aligned with EU fundamental values. TRAIL activities are often

[7] https://www.doriangray-horizon.eu/ - Last visited 30 May 2025

anchored in the Z-Inspection® methodology. Looking forward, the lab is committed to refining its methodologies, expanding cross-sector collaborations, and enhancing AI governance to foster transparency, accountability, and public trust.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 2.5 and Grammarly to check grammar and spelling. After using these tool the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. V. Zicari, J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, G. Roig, N. Stürtz, K. Tolle, J. J. Tithi, I. van Halem, M. Westerlund, Z-inspection®: A process to assess trustworthy ai, IEEE Transactions on Technology and Society 2 (2021) 83–97. doi:10.1109/TTS.2021.3066209.

[2] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, D. Farina, Bs-net: Learning covid-19 pneumonia severity on a large chest x-ray dataset, Medical Image Analysis 71 (2021). doi:10.1016/j.media.2021.102046.

[3] H. Allahabadi, J. Amann, I. Balot, A. Beretta, C. Binkley, J. Bozenhard, F. Bruneault, J. Brusseau, S. Candemir, L. A. Cappellini, S. Chakraborty, N. Cherciu, C. Cociancig, M. Coffee, I. Ek, L. Espinosa-Leal, D. Farina, G. Fieux-Castagnet, T. Frauenfelder, A. Gallucci, G. Giuliani, A. Golda, I. van Halem, E. Hildt, S. Holm, G. Kararigas, S. A. Krier, U. Kühne, F. Lizzi, V. I. Madai, A. F. Markus, S. Masis, E. W. Mathez, F. Mureddu, E. Neri, W. Osika, M. Ozols, C. Panigutti, B. Parent, F. Pratesi, P. A. Moreno-Sánchez, G. Sartor, M. Savardi, A. Signoroni, H.-M. Sormunen, A. Spezzatti, A. Srivastava, A. F. Stephansen, L. B. Theng, J. J. Tithi, J. Tuominen, S. Umbrello, F. Vaccher, D. Vetter, M. Westerlund, R. Wurth, R. V. Zicari, Assessing trustworthy ai in times of covid-19: Deep learning for predicting a multiregional score conveying the degree of lung compromise in covid-19 patients, IEEE Transactions on Technology and Society 3 (2022) 272–289.

[4] M. Savardi, A. Signoroni, S. Benini, F. Vaccher, M. Alberti, P. Ciolli, N. Di Meo, T. Falcone, M. Ramanzin, B. Romano, F. Sozzi, D. Farina, Upskilling or deskilling? measurable role of an ai-supported training for radiology residents: a lesson from the pandemic, Insights into Imaging 16 (2025). doi:10.1186/s13244-024-01893-4.