TRUST-AI 2025 – POSITION PAPER REPORT

Position papers presented at the European Workshop on Trustworthy AI, October 25-26, 2025

Report editors:

Asbjørn Følstad, Dimitris Apostolou, Steve Taylor, Andrea Palumbo, Eleni Tsalapati, Giannis Stamatellos, Rosario Catelli

---

TRUST-AI was organized as part of the European Conference on Artificial Intelligence, ECAI – 2025.

Preface to the Position Paper Report of TRUST-AI 2025 – the European Workshop on Trustworthy AI

Introduction

In response to the need for research and collaboration on trustworthy AI, TRUST-AI 2025 – the European Workshop on Trustworthy AI was organized as part of the European Conference on Artificial Intelligence, ECAI – 2025, in Bologna, October 25-26, 2025.

This report includes the position papers accepted for presentation at TRUST-AI 2025. Position papers is a paper format for shorter contributions that present specific positions or open questions in need of reflection or discussion. The relevance and quality of the accepted position papers suggests that they may be of general interest also for a broader audience. We were therefore motivated to collate these in the present report as a resource for researchers and practitioners with an interest in trustworthy AI.

The position papers were submitted in response to the workshop call for papers. This call received substantial interest with 60 papers submitted overall, of these 10 position papers. Following an initial screening process, each of the submitted papers were reviewed by three independent reviewers. Final decisions on paper acceptance were made by the workshop organizers.

The accepted position papers cover a range of topics pertaining to trustworthy AI. Mateescu discusses a lifecycle approach to mitigate epistemic risk in adaptive AI systems. Käsbohrer et al. address how trust in AI applications can be fostered through counterfactual explanations and causal reasoning. Darnell et al. discussed trust and trustworthiness of generative AI tools. Di Scala et al. reflects on the potential gap between functional and normative aspects of trustworthy AI. Sandulu et al. present an AI debugger for practical assistance during trustworthy AI development. Jazayeri et al. propose a risk index for evaluating AI system evaluation in terms of compliance with trustworthy AI principles and deployment risk factors. Mala et al. present a trustworthy AI design case in the form of a generative chatbot for public administration. And, finally, Villalobos-Quesada discuss how labelling may be an approach towards strengthening trustworthy AI.

This position paper report is provided in addition to the main workshop proceedings published by CEUR Workshop Proceedings, where the latter includes the majority of papers accepted for the workshop. Hence, for a full overview of the workshop outcomes please consider the main workshop proceedings (https://ceur-ws.org/Vol-4132/) as well as this report.

Workshop organizers

- Asbjørn Følstad, SINTEF, Norway
- Dimitris Apostolou, ICCS & University of Piraeus, Greece
- Steve Taylor, University of Southampton, UK
- Andrea Palumbo, KU Leuven, Belgium
- Eleni Tsalapati, ATC, Greece
- Giannis Stamatellos, Institute of Philosophy & Technology, Greece
- Rosario Catelli, Engineering, Italy

Acknowledgments

The workshop was supported by EU Horizon Europe, HORIZON-CL4-2022-HUMAN-02-01, under the project THEMIS 5.0 (grant agreement No. 101121042).

The paper template applied for the position papers in this report is based on *A better way to format your document for CEUR-WS,* by Kulyabov, D. S., Tiddi, I., & Jesusfield, M. https://www.overleaf.com/latex/templates/template-for-submissions-to-ceur-workshop-proceedings-ceur-ws-dot-org/wqyfdgftmcfw

Content

This position paper report includes the following contributions:

Human-Centered Risk Governance for Adaptive AI: Why Educational Requirements Belong in Trustworthy AI Frameworks

Alexandru Mateescu[1]

[1]Université Paris 1 Panthéon-Sorbonne, École Doctorale 280, Department of Philosophy, Paris, France

Abstract

Adaptive AI systems — those that evolve through user interaction — create distinctive epistemic risks: users may lose track of how the system changes, why outputs are produced, or how their actions shape future recommendations. These risks go beyond standard concerns such as fairness or robustness and are not addressed by existing safeguards. We argue that trust in such systems should be understood as a relational process involving both users and designers. From this perspective, we introduce educational requirements: lifecycle design commitments embedded throughout the AI system to promote user intelligibility, reflection, and control. This approach aligns with the Human-Centered AI perspective, which emphasizes combining high levels of automation with equally high levels of human control to ensure reliable, safe, and trustworthy systems [1].

Beyond delivering information, these requirements must also ensure that educational content is received and acknowledged. Emerging content management technologies — including blockchain-based approaches — could soon enable the tracking of distribution and uptake of key educational snippets, providing a verifiable record of user engagement and accountability.

Grounded in regulatory frameworks such as the AI Act and the Digital Services Act, these requirements offer a proportionate and forward-looking way to mitigate epistemic risks and support human-centered AI governance.

1. Introduction

Adaptive AI systems — those that evolve through user interaction — are now a central concern for regulators. Frameworks such as the NIST AI Risk Management Framework [2], ENISA guidelines [3], and the European AI Act [4] emphasize the need for oversight of systems that continue to learn after deployment.

Adaptivity introduces distinctive risks. When AI systems evolve by responding to user behavior — as in many recommender contexts — they also shape that behavior in return. This feedback loop creates epistemic risks: users may lose track of how the system changes, why outputs are produced, or how their actions influence future recommendations.

These dynamics challenge conventional models of trustworthiness. Existing safeguards such as fairness, explainability, and robustness are necessary but insufficient. Trust must instead be understood as a relational process co-constructed by users and designers over time.

We propose educational requirements: design commitments embedded throughout the AI lifecycle to promote user intelligibility, reflection, and control. Grounded in emerging regulations such as the AI Act and the Digital Services Act, these requirements offer a proportionate way to mitigate epistemic risks and support more accountable, human-centered AI governance.

2. Adaptivity and Epistemic Risk

Adaptive AI systems do not follow fixed rules: they continuously adjust their outputs based on user interactions and behavioral patterns. While this improves personalization, it also creates distinctive epistemic risks. Users may lose track of how the system evolves, why certain outputs are produced, or how their actions shape future recommendations.

TRUST-AI Workshop, ECAI 2025, Bologna, Italy — October 25, 2025

alexandru.mateescu@etu.univ-paris1.fr (A. Mateescu)

These risks differ from conventional concerns such as fairness or robustness. They undermine the user's ability to interpret and influence a system whose logic shifts over time. The problem is most acute when users are treated merely as variables in optimization, rather than as epistemic agents capable of making sense of the process.

This opacity can foster harms such as biased beliefs, dependency, or disengagement. It also challenges regulatory goals like transparency, which presume that users can understand and contest significant automated decisions [4].

Addressing these risks requires mechanisms that go beyond technical transparency and documentation. Prior work in VIS4ML highlights that many interactive ML systems remain limited to case studies and lack verifiable evidence of real-world impact, underscoring the need for governance approaches that integrate verification and accountability [5]

3. Framing Trust as a Relational Epistemic Process

Standard approaches to trustworthy AI focus on system attributes such as fairness, accuracy, robustness, or explainability. These are essential but reflect an internalist view, where trustworthiness is judged solely by the system's properties. In adaptive contexts, this view is insufficient.

Trust must instead be seen as relational: an evolving process between users and systems that themselves evolve through interaction. Adaptive AI not only learns from users but also shapes their behavior. Trust depends on the user's ongoing capacity to interpret, evaluate, and contest this dynamic.

Promoting trustworthiness therefore requires supporting the user's epistemic role, not just improving algorithms. This involves making adaptation mechanisms visible and giving users tools to track and influence outcomes.

This perspective aligns with regulatory goals in the AI Act [4] and the Digital Services Act [6], which emphasize transparency and oversight. Implementing these principles requires more than documentation: it calls for designs that embed users in the learning loop as interpreters and co-regulators, not merely data points.

The next section introduces educational requirements as a practical way to operationalize this relational approach and mitigate epistemic risks.

4. Educational Requirements as Governance Mechanisms

Educational requirements are system-level design commitments that make adaptivity intelligible to users and actionable for designers. Unlike external documentation or user training, they are embedded directly into the interface, feedback loops, and lifecycle processes of an adaptive AI system. Their purpose is to ensure that as the system evolves, users remain active epistemic agents rather than passive data sources.

We distinguish between two roles:

1. **End users** – whose interactions generate the data that drive adaptation.
2. **Designers and developers** – who define the mechanisms of adaptation and must anticipate the epistemic risks they create.

Failures arise when these roles are misaligned — for instance, when a recommender system is optimized to maximize engagement while users seek informed decisions [7]. Educational requirements help bridge this gap by embedding mutual intelligibility into the design process.

Examples from recent work illustrate how educational interventions can be embedded directly into recommender systems to foster user understanding and agency [8, 9].

Explanations alone are insufficient: research shows they reduce overreliance only for relatively easy decisions where users can already verify the correct answer, but fail to support decision-making in harder cases [10].Concretely, these requirements can take several forms:

- **Traceability indicators** – showing how current outputs relate to past interactions and how future recommendations may shift as a result.
- **Dynamic transparency prompts** – surfacing which signals the system is prioritizing, reducing the "black-box" effect of invisible adaptation.
- **Feedback channels** – allowing users to contest or redirect system learning when outputs misalign with their goals.
- **Verification of educational uptake** – mechanisms that assess whether key educational snippets have been delivered and processed by the user. Emerging content management capabilities, including blockchain-based approaches, make it possible to track the distribution and acknowledgment of such snippets while preserving accountability and user rights.

Implementing these mechanisms raises practical challenges. Excessive detail risks overwhelming users, while oversimplification fosters false confidence. Commercial incentives may also conflict with the goal of user education, as platforms often benefit from opacity or frictionless engagement. Addressing these tensions requires deliberate trade-offs, balancing intelligibility, usability, and business goals.

Because they can be integrated at design, development, and testing stages, educational requirements are compatible with lifecycle governance. They operationalize human-centered oversight, ensuring that epistemic risks are treated not as side effects but as core governance targets, alongside explainability and fairness.

5. Regulatory Alignment and Proportionality

Educational requirements are not only ethically motivated — they also align with emerging legal frameworks. The Digital Services Act (DSA) introduces the concept of systemic risks, including threats to civic discourse and individual agency, and emphasizes proportional mitigation obligations [6]. Similarly, the AI Act calls for governance measures tailored to the context and risk level of specific systems [4]. Adaptive AI systems, especially recommender systems, often fall into gray zones: they shape user behavior in subtle ways that may not trigger strict legal thresholds but still cause epistemic harm. Educational requirements provide provide a balanced alternative to both heavy-handed intervention and passive transparency. They allow designers to address these risks proactively without limiting system capabilities or imposing intrusive oversight.

Because they can be integrated at the stages of design, development, and testing, educational requirements fit naturally into compliance checklists, certification schemes, and audits as the AI Act moves toward enforcement. By making the user's epistemic position a measurable aspect of trustworthiness, they extend the principle of proportionality into the cognitive domain, ensuring that adaptive systems support not only legal compliance but also the user's right to understand and shape algorithmic processes over time. The principle of proportionality, central to both the AI Act and the DSA, ensures that governance measures are calibrated to the level of risk: more stringent obligations for systems with greater potential harm, and lighter measures for those with minimal impact.

6. Conclusion

Adaptive AI systems create new forms of epistemic risk that conventional safeguards such as fairness, explainability, or robustness cannot fully address. These risks emerge from feedback loops where users both influence and are influenced by system behavior, often without clear visibility into this process.

We have argued that trust in adaptive AI should be understood as a relational process involving both users and designers. From this perspective, we proposed educational requirements: lifecycle design commitments that embed intelligibility and user agency directly into adaptive systems.

Crucially, these requirements must go beyond simply presenting information. Systems should be able to assess whether educational snippets have been delivered and engaged with, ensuring that users are not only exposed to content but also empowered to act on it. Emerging content management capabilities,

potentially supported by blockchain technologies, could soon allow for verifiable records of distribution and acknowledgment. This would strengthen both accountability and regulatory compliance while preserving user rights.

By aligning with the AI Act and Digital Services Act, educational requirements become a proportionate governance mechanism, bridging the gap between formal legal compliance and meaningful user oversight. We invite researchers, designers, and regulators to explore and refine these requirements, and to collaborate on developing the technological foundations — including content management and blockchain-based tracking — needed to verify the delivery and acknowledgment of educational snippets. Working together, we can ensure that adaptive AI systems remain transparent, intelligible, and accountable as they evolve.

References

[1] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, arXiv preprint arXiv:2002.04087 (2020). URL: https://arxiv.org/abs/2002.04087.

[2] NIST, Ai risk management framework (ai rmf 1.0), https://www.nist.gov/itl/ai-risk-management-framework, 2023. Accessed: 2025-09-20.

[3] ENISA, Artificial intelligence cybersecurity challenges, https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges, 2020. Accessed: 2025-09-20.

[4] E. Commission, Artificial intelligence act – final political agreement, https://artificial-intelligence-act.eu/, 2024. Accessed: 2025-09-20.

[5] H. Subramonyam, J. Hullman, Are we closing the loop yet? gaps in the generalizability of vis4ml research, IEEE Transactions on Visualization and Computer Graphics 29 (2023) 1–15. doi:10.1109/TVCG.2023.3243440.

[6] E. Commission, Digital services act, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package, 2022. Accessed: 2025-09-20.

[7] N. Seaver, Captivating algorithms: Recommender systems as traps, Journal of Material Culture 28 (2023) 3–21. doi:10.1177/13591835221081874.

[8] L. Alves, R. Cardoso, R. B. C. Prudêncio, Digital nudges for recommender systems: Managing exploration and exploitation trade-offs, Information Processing & Management 61 (2024) 103269. doi:10.1016/j.ipm.2023.103269.

[9] V. Lomonaco, D. Maltoni, et al., Game-based education to address filter bubbles and echo chambers, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2022, pp. 905–917. doi:10.1145/3531146.3534634.

[10] Z. T. Zhang, F. Buchner, Y. Liu, A. Butz, You can only verify when you know the answer: Feature-based explanations reduce overreliance on ai for easy decisions, but not for hard ones, in: Proceedings of Mensch und Computer 2024 (MuC '24), 2024, pp. 1–15. URL: https://doi.org/10.1145/3670653.3670660. doi:10.1145/3670653.3670660.

The Right to Distrust: Designing Clinical AI for Robust Comparison

Cornelia C. Käsbohrer[1], Tim Barz-Cech[2] and Lili Jiang[1]

[1] *Umeå University, Sweden*
[2] *HMS Analytical Software GmbH, Germany*

Abstract

Many clinicians remain hesitant to rely on AI systems in high-stakes decision-making, particularly when models are opaque or poorly aligned with clinical reasoning. A common approach to achieve this often relies on visual add-ons such as saliency maps in the context of medical imaging. However, we argue that such efforts fall short, since saliency maps are correlational, difficult to interpret, and disconnected from the causal logic that physicians apply when deciding whether to biopsy or treat a lesion. We call for a shift in focus. Rather than trying to persuade physicians to trust AI, we should consider which forms of distrust are justified. Specifically, we propose that causal, counterfactual explanations presented via familiar, image-based interfaces provide a more robust basis for justified reliance. We present a design and evaluation plan for an interactive cancer imaging viewer that facilitates counterfactual exploration and causal reasoning. We also propose practical methods for measuring its impact on physician trust.

1. Introduction

Asking *"How do we make physicians trust AI?"* may seem like the natural place to begin, but it risks overlooking a deeper issue. As physicians are legally and ethically responsible for every treatment decision they make, skepticism toward opaque automation is a professional virtue, rather than a barrier to innovation. Despite the increasing integration of AI into diagnostic workflows, many clinicians remain hesitant to rely on such systems in high-stakes decision making, particularly when the underlying models are opaque or not aligned with clinical reasoning [1]. As Kraemer et al. [2] argue, ethical decision making is essential when algorithms are used in high-stakes situations. Physicians must critically assess systems whose outputs have significant moral implications. Decades of automation research demonstrate that users only rely on systems they can comprehend and manipulate [3, 4]. However, most medical imaging models reveal little more than a probability score and a heat map of "focused" pixels. Radiologists are expected to pose explanatory questions, such as identifying the cause of the malignancy score and considering the effect of alterations to the image or the patient.

2. Background and Related Work

Correlation does not imply causation. Saliency methods highlight where the network is focused, but not why a particular region is associated with malignancy or how it could be altered [5, 6, 7]. There is a lack of clinical rigor and actionability. Heatmaps rarely align with established

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.
cornelia.kaesbohrer@cs.umu.se (C. C. Käsbohrer)
0009-0004-0702-6227 (C. C. Käsbohrer); 0000-0001-8688-2419 (T. Barz-Cech); 0000-0002-7788-3986 (L. Jiang)

radiological descriptors (e.g., spiculation or necrosis) and cannot be included in staging reports or tumour board discussions. Moreover, there is limited empirical evidence that these tools support calibrated trust, particularly in high-stakes clinical tasks [8]. User studies show mixed or negligible gains in diagnostic accuracy and confidence when saliency maps are added to AI outputs. Although rarely emphasized in evaluation studies, counterfactual[1] reasoning tasks could help address the disconnect between model salience and clinical reasoning by revealing how feature changes affect predictions.

While counterfactual image generation has gained interest in general computer vision [9, 10, 11], to the best of our knowledge, no prior work has demonstrated its effectiveness in clinical settings where predictions rely on subtle visual patterns, expert interpretation, and high-stakes decisions. Medical images differ fundamentally from natural images: they exhibit less redundancy, operate under stricter anatomical constraints, and require interpretability grounded in physiological realism [12]. Moreover, the design of explanatory interfaces for medical experts must go beyond technical accuracy by incorporating domain-specific language and workflows [13]. A recent systematic review of 68 transparency-focused medical imaging studies found that only three involved end-user evaluation, and none incorporated interactive causal or contrastive interfaces aligned with clinical reasoning processes [14]. To our knowledge, and as a recent review supports, no published system enables radiologists to explore slight image variations that produce different model outputs, despite the potential of such comparisons to reveal decision boundaries [14]. Such an image would not only support clinical reasoning but also serve as a robustness test, revealing how close a prediction is to decision boundaries. Current imaging-AI systems lack the ability to compare visually similar examples with opposing outcomes, which is critical for fostering physician trust.

3. Proposed Idea

We argue that trustworthy AI-powered imaging should be designed to facilitate meaningful comparisons rather than persuasive explanations. Saliency maps often prompt speculative interpretations because they highlight regions without providing a contextual explanation. However, physicians engage in structured exploratory reasoning, such as considering how small, clinically plausible changes might alter a diagnosis. Supporting this kind of reasoning requires explanations that go beyond visual focus to capture causal or counterfactual relationships. Such contrastive insight, delivered through visually plausible counterfactual examples, could provide a clearer understanding of a model's robustness, limitations, and reasoning boundaries. This changes the goal. Rather than expecting physicians to apply causal reasoning to model outputs, explanation systems should externalize that reasoning. Practitioners can then scrutinize predictions by making minimal, semantically coherent changes. This approach reduces cognitive burden and aligns more closely with clinicians' decision-making needs in high-stakes contexts. Thus, we align our explanations with their usual mode of reasoning to lower the mental load required to interact with our system.

Moreover, we draw on design principles tailored for domain experts, particularly the guidance to use familiar terminology and support learning through progressive refinement [13]. For radiologists, this means that explanations should be based on the language of structured reporting (e.g., margin, enhancement, and spiculation) and be designed to enable the step-by-

[1] Given a trained imaging model f and a case x, a counterfactual explanation is a near-identical, clinically-plausible example x' that differs from x by a minimal, semantically meaningful change and causes f to change in a specified direction (e.g., to cross the decision boundary).

step exploration of model behaviour. This is implemented in our proposed interface (see Table 1), which uses a PACS (Picture Archiving and Communication System) viewer [15] that is already familiar to radiologists, enabling them to toggle between the original image and a counterfactual image. By grounding explanations on established tools and terminology and enabling progressive interaction, our aim is to promote transparency without causing cognitive overload.

Table 1
Proposed interface elements for trust-enhancing counterfactual comparison

Familiar element to clinicians	Explanation extension	Interaction
PACS slice viewer [15]	Toggle: original ↔ nearest image with flipped prediction	Scroll/zoom as usual
Side-by-side display	Shows prediction, confidence, and clinical similarity score	Visual + numeric comparison
Risk slider	Indicates how prediction confidence changes near decision boundary	Interactive movement through latent space
Evidence panel	Retrieve [16] or synthesize [17, 18, 19] similar real cases with same vs. opposite labels	Click-through to full pathology report

Prototype development Create a pipeline that retrieves [16] or synthesizes [17, 18, 19] clinically realistic counterfactuals. These counterfactuals should be minimally and mechanistically different from the current case, but lead to an opposite prediction. In this context, similarity is not just mathematical, but also based on plausible medical variation, as required by counterfactual reasoning frameworks [20]. This does not require full image editing, but could leverage latent space navigation within a diffusion or encoder-decoder model that is constrained to maintain anatomical plausibility. We prioritise the retrieval of clinically similar cases with opposite labels. When synthesis is required, edits are constrained by segmentation-preserving losses, descriptor-stability regularisers and acceptance thresholds (ε_{IoU}, descriptor bands), which were tuned in a pilot study with radiologists.

Interface prototype Develop a PACS-style viewer that displays the original image alongside a minimally altered version with a flipped prediction, as well as the prediction confidence and semantic similarity. The interface should incorporate design principles for domain experts, such as structured vocabulary and layered exploration.

User study with radiologists (n = 10) The evaluation will compare three interface variants in a within-subject design:

(A) a baseline black-box AI interface with no explanation,

(B) an interface that overlays standard saliency maps on the original image, and

(C) a novel interface that shows counterfactual comparisons-images that are visually similar but yield opposite model predictions.

Each participant will interact with all three conditions, allowing us to measure how each explanation format affects diagnostic confidence, trust calibration, and cognitive workload.

Metrics The metrics we will use include diagnostic confidence [8], perceived robustness [21], NASA-TLX [22], and System Causability Scale [23].

Limitations As the sample size is small, the results of the pilot study may be underpowered and sensitive to participant composition. System Causability Scale captures perceived causability rather than clinical safety or utility and must be interpreted alongside behavioral outcomes. Data access and privacy constraints limit availability of pathology-verified,

4. Open Questions for Discussion

Important open questions remain regarding the feasibility and implementation of our proposed approach. The first question is what level of realism and similarity is required for physicians to accept a counterfactual image as clinically useful and trustworthy. Next, we ask whether techniques such as latent-space traversal or retrieval of real-world analogues are sufficient to generate flipped-prediction examples, or if full generative modeling is necessary. Furthermore, how can we constrain these explanations to maintain clinical plausibility and anatomical correctness, particularly given the sensitivity of interpreting medical images? Finally, how can we evaluate whether such contrastive explanations increase physicians' sense of being informed and confident in their decision-making rather than merely persuading them through visual cues?

5. Conclusion

Rather than creating systems that demand trust, we suggest developing systems that empower physicians to question and test predictions. Our novelty is the *integration and evaluation* of a PACS-style counterfactual comparison workflow, not new generators or metrics. In clinical image interpretation, viewing a near-identical case with a flipped model prediction may reveal model robustness more effectively than any heatmap. This shift from explanation to contrastive evaluation offers a novel, clinically aligned approach to foster meaningful, evidence-based trust in AI.

Acknowledgments

This paper is a result of the discussion at ELLIIT Focus Period "Visualization-Empowered Human-in-the-Loop Artificial Intelligence" at Linköping University, Campus Norrköping, Sweden, Spring 2025.
This research is funded by the European Union (COMFORT-Computational Models FOR patienT stratification in urologic cancers – Creating robust and trustworthy multimodal AI for health care), 101079894.

References

[1] J. R. Geis, A. P. Brady, C. C. Wu, J. Spencer, E. Ranschaert, J. L. Jaremko, S. G. Langer, A. Borondy Kitts, J. Birch, W. F. Shields, et al., Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement, Radiology 293 (2019) 436–440.
[2] F. Kraemer, K. Van Overveld, M. Peterson, Is there an ethics of algorithms?, Ethics and Information Technology 13 (2011) 251–260.
[3] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International Journal of Human-Computer Studies 58 (2003) 697–718.

[4] E. Glikson, A. W. Woolley, Human trust in artificial intelligence: Review of empirical research, Academy of management annals 14 (2020) 627–660.
[5] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un)reliability of saliency methods, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (2019) 267–280.
[6] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems, volume 31, 2018.
[7] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, P. Kessel, Explanations can be manipulated and geometry is to blame, in: Advances in Neural Information Processing Systems, volume 32, 2019.
[8] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al., Human–computer collaboration for skin cancer recognition, Nature Medicine 26 (2020) 1229–1234.
[9] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, Counterfactual visual explanations, in: International Conference on Machine Learning, PMLR, 2019, pp. 2376–2384.
[10] R. Bischof, F. Scheidegger, M. A. Kraus, A. C. I. Malossi, Counterfactual image generation for adversarially robust and interpretable classifiers, arXiv preprint arXiv:2310.00761 (2023).
[11] F. Böttger, T. Cech, W. Scheibel, J. Döllner, Visual counterfactual explanations using semantic part locations, in: Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR, 2023, pp. 63–74.
[12] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.
[13] Y. L. Wong, K. Madhavan, N. Elmqvist, Towards characterizing domain experts as a user group, in: 2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV), IEEE, 2018, pp. 1–10.
[14] H. Chen, C. Gomez, C.-M. Huang, M. Unberath, Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review, NPJ Digital Medicine 5 (2022) 156.
[15] R. H. Choplin, J. Boehme 2nd, C. D. Maynard, Picture archiving and communication systems: an overview., Radiographics 12 (1992) 127–129.
[16] A. Gottlieb, G. Y. Stein, E. Ruppin, R. B. Altman, R. Sharan, A method for inferring medical diagnoses from patient similarities, BMC Medicine 11 (2013) 194.
[17] Y. Gu, J. Yang, N. Usuyama, C. Li, S. Zhang, M. P. Lungren, J. Gao, H. Poon, Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys, arXiv preprint arXiv:2310.10765 (2023).
[18] M. B. Alaya, D. M. Lang, B. Wiestler, J. A. Schnabel, C. I. Bercea, Mededit: Counterfactual diffusion-based image editing on brain mri, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer, 2024, pp. 167–176.
[19] M. Atad, D. Schinz, H. Moeller, R. Graf, B. Wiestler, D. Rueckert, N. Navab, J. S. Kirschke, M. Keicher, et al., Counterfactual explanations for medical image classification and regression using diffusion autoencoder, Machine Learning for Biomedical Imaging 2 (2024) 2103–2125.
[20] G. Keil, Making causal counterfactuals more singular, and more appropriate for use in law, Benedikt Kahmen/Markus Stepanians (Hg.), Causation and Responsibility. Critical Essays, Berlin/Boston (2013) 157–189.
[21] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv

preprint arXiv:1702.08608 (2017).
[22] S. G. Hart, L. E. Staveland, Development of nasa-tlx (task load index): Results of empirical and theoretical research, in: Advances in Psychology, volume 52, Elsevier, 1988, pp. 139–183.
[23] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.

Trust as an Outcome of Trustworthy AI: A Case for Increasing Research of Trust of Agentic AI Tools

Elizabeth Darnell[1,*], Emma Murphy[1] and Dympna O'Sullivan[1]

[1]*Technological University Dublin, Dublin, Ireland*

Abstract

Generative artificial intelligence tools are being widely adopted and integrated across industries. In the European context, there is a greater focus on trustworthy AI due to the EU AI Act. Trust is an outcome of effective trustworthy AI systems, however, there is a lack of research on trust of generative AI tools due to the rapid evolution of the tools. Generative AI tools are disrupting the workforce, especially in computing fields but there has not been significant research on evaluating trust of these tools. This paper presents a protocol to measure and understand trust in the context of agentic coding assistants in the workplace.

Keywords

Trust, Trustworthy AI, Human-AI Trust, Human-AI Interaction, Agentic Coding Assistants, Agentic AI Tools

1. Introduction

Since the public launch of ChatGPT in late 2022, there has been extensive reporting about the potential and current impact of AI on the workforce [1, 2, 3]. In early 2025, companies like Google and Microsoft have reported that generative AI-powered tools are currently generating at least a quarter of their new code and anticipating more growth indicating that these types of companies are widely adopting agentic coding tools and practices [4, 5]. Simultaneously, trustworthy AI (TAI), is at the forefront of many discussions about generative AI systems and tools, especially in the EU, due to the EU AI Act positioning TAI principles as a key component of the legislation [6, 7]. Given the fast emergence and rapidly changing landscape of these tools, there has not yet been a significant amount of attention placed on understanding and measuring trust despite trust being an outcome of effective TAI systems [8]. In this paper we propose that there should be a greater research focus on trust of generative AI powered agentic coding assistants due to their growing influence on software development practices.

2. Background

2.1. Defining and Understanding Trust

Trust is a difficult dimension to define, measure, and understand and lacks consistent methodology in research across disciplines [8, 9, 10, 11, 12, 13]. Despite there not being a definitive definition of trust, the following is a frequently used definition and the definition that this work uses is "trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behaviour of another" [13, p. 395]. This definition is appropriate for this research as it highlights the individual nature of trust while introducing the requirement of vulnerability to develop trust. In the context of adoption and use of agentic coding assistants the positive expectations is the mapped to the accuracy and quality of the generated code while the acceptance of vulnerability is mapped to the understanding that individual's that utilise the generated code are ultimately responsible for the generated code. Trust is contextual and changing the context in which an individual is trusting,

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

elizabeth.g.darnell@mytudublin.ie (E. Darnell)

0009-0007-9204-2492 (E. Darnell)

can change the level of trust [8, 14, 15, 16]. There is also the question of whether trust and distrust are a single dimension on opposite sides of a continuum or if trust and distrust are separate but related constructs [9, 8, 17, 10, 12]. Recent research has begun to move towards the understanding of trust and distrust as distinct constructs [9, 10, 8]. While there has been some focus on trust of AI systems in e-commerce and medicine, there has not been as much exploration of trust in software development fields including agentic coding assistants [18, 19, 20, 21, 22]. Agentic coding assistants provide a practical use case to explore and measure trust, especially given that the outputs of these tools are more quantifiable than other use cases.

2.2. Measuring Trust

Scales, both custom and standardised, and qualitative methods are most commonly used to understand and measure trust [10, 9]. Custom scales are frequently developed and utilised for the specific context being studied which hinders reproducability due to the highly contextual nature of trust [10, 8, 9]. While limited, there are a small number of standardised and validated scales that have been applied in this research area, for example the Trust in Automation scale, [23] which has become a more frequently utilised trust scale which allows for more standardisation and comparison across contexts and studies [11, 9, 8]. The Trust in Automation scale initially concluded that trust and distrust are on the same continuum [23]. Recent research indicates that the Trust in Automation scale can also be applied successfully with the understanding of trust and distrust as distinct concepts [9, 10, 8, 12].
Interviews are heavily relied upon as a qualitative method to understand and measure trust as it allows for insights on how individuals conceive of trust while remaining grounded in the specific context [10, 24, 25, 16]. Trust is known to impact the adoption of and reliance on new technologies with use of a tool or technology [15, 26, 8]. While reliance and use of a tool are sometimes used as a proxy for trust, trust cannot occur without vulnerability but end-users can use and rely on tools in environments without actual or perceived vulnerability [8]. So, the use of reliance and general use as a proxy of trust should be used cautiously [8]. Given the necessity of vulnerability to be present for trust to occur it is important to bring up the concept of appropriate trust or trust calibration. The concept of trust calibration recognizes that trust is not static and ideally, should be related to the ability of the tool itself [12, 26]. The lack of consistent, reliable, and validated methods of measurement hinders our ability to understand and measure trust in clear and reproducible ways.

3. Discussion

Given that trust is an outcome of TAI, there needs to be greater research attention to understanding and measuring trust. Understanding the trust of agentic coding assistants is important due to the widespread adoption, their impact on overall codebase quality, and the impact on the labour-force [27]. The use of these tools has broad impact at these companies and evaluating their trust is complex as it could be evaluated based upon the tool, the process, and the resulting generated code. Additionally, there are traits of the end-user that might impact the trust of these tools, such as their technical expertise, age, or gender. However, there has not been significant research into the impact end-user traits on trust in general, and especially not in AI contexts. This is important due to the potential impact on the workforce and adoption depending on how individuals trust these agentic coding assistants and if there is a relationship to level of trust and different traits which could create uneven adoption and reliance. We are currently running a study utilising a novel mixed-method protocol to evaluate trust of agentic coding assistants in the workplace. The protocol has three phases comprising of a questionnaire regarding the participant's current generative AI use at work, the observation of an open-exploration task with an agentic coding assistant followed by the completion of Trust in Automation scale, and concluding with a semi-structured interview. The protocol as such contains both attitudinal and behavioural measures of trust allowing for a more complete understanding of trust. The participants are experts in software development or have significant experience with coding. We hope that this protocol contributes to

expanding this body of knowledge and is able to be used across AI contexts improving the understanding of trust of generative AI tools.

Acknowledgments

This work was supported by the TU Dublin ARISE (Amplify Research & Innovation Supporting Enterprise) programme, funded under the TU RISE scheme, and co-financed by the Government of Ireland and the European Union through the ERDF Southern, Eastern & Midland Regional Programme 2021–2027 and the Northern & Western Regional Programme 2021–2027.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. McNeilly, Will Generative AI Disproportionately Affect the Jobs of Women? , Technical Report, Kenan Institute of Private Enterprise UNC Kenan-Flagler Business School, 2023. URL: https://kenaninstitute.unc.edu/kenan-insight/will-generative-ai-disproportionately-affect-the-jobs-of-women/.

[2] K. Roose, This A.I. Forecast Predicts Storms Ahead, 2025. URL: https://www.nytimes.com/2025/04/03/technology/ai-futures-project-ai-2027.html.

[3] O. Dmitracova, 41% of companies worldwide plan to reduce workforces by 2030 due to AI, 2025. URL: https://edition.cnn.com/2025/01/08/business/ai-job-losses-by-2030-intl/index.html.

[4] J. Novet, J. Vanian, Satya Nadella says as much as 30% of Microsoft code is written by AI, 2025. URL: https://www.cnbc.com/2025/04/29/satya-nadella-says-as-much-as-30percent-of-microsoft-code-is-written-by-ai.html.

[5] M. Temkin, M. Zeff, Why OpenAI wanted to buy Cursor but opted for the fast-growing Windsurf, 2025. URL: https://techcrunch.com/2025/04/22/why-openai-wanted-to-buy-cursor-but-opted-for-the-fast-growing-windsurf/.

[6] P. Office of the European Union L, L. Luxembourg, Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj.

[7] L. McCormack, M. Bendechache, A Comprehensive Survey and Classification of Evaluation Criteria for Trustworthy Artificial Intelligence (2024). URL: http://arxiv.org/abs/2410.17281.

[8] R. Visser, T. M. Peters, I. Scharlau, B. Hammer, S. Thill, Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation, Cognitive Systems Research (2025) 1–9. URL: https://doi.org/10.1007/978-3-. doi:10.1007/978-3.

[9] N. Scharowski, S. A. C. Perrig, N. von Felten, L. F. Aeschbach, K. Opwis, P. Wintersberger, F. Brühlmann, To Trust or Distrust AI: A Questionnaire Validation Study, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2025, pp. 361–374. URL: https://dl.acm.org/doi/10.1145/3715275.3732025. doi:10.1145/3715275.3732025.

[10] O. Vereschak, G. Bailly, B. Caramiaux, How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies, Proceedings of the ACM on Human-Computer Interaction 5 (2021). doi:10.1145/3476068.

[11] R. Chen, R. Wang, N. Sadeh, F. Fang, Missing Pieces: How Do Designs that Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2025, pp. 790–816. URL: https://dl.acm.org/doi/10.1145/3715275.3732050. doi:10.1145/3715275.3732050.

[12] J. D. Lee, K. A. See, Trust in Automation: Designing for Appropriate Reliance, Human Factors 46 (2004) 50–80.
[13] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer, Not so different after all: a cross-discipline view of trust, Academy of Management Review 23 (1998) 393–404.
[14] O. G. McKinley, S. Pandey, A. Ottley, Trustworthy by Design: The Viewer's Perspective on Trust in Data Visualization, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2025, pp. 1–17. URL: https://dl.acm.org/doi/10.1145/3706598.3713824. doi:10.1145/3706598.3713824.
[15] R. Yang, S. Wibowo, User trust in artificial intelligence: A comprehensive conceptual framework, Electronic Markets 32 (2022) 2053–2077. doi:10.1007/s12525-022-00592-6.
[16] R. J. Lewicki, C. Brinsfield, Measuring trust beliefs and behaviours, in: F. Lyon, G. Möllering, M. Saunders (Eds.), Handbook of Research Methods on Trust, Edward Elgar Publishing Limited, Cheltenham, 2012, pp. 29–39.
[17] S. Göbel, R. Lämmel, Model-Based Trust Analysis of LLM Conversations, in: Proceedings: MODELS 2024 - ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, Association for Computing Machinery, Inc, 2024, pp. 602–610. doi:10.1145/3652620.3687809.
[18] D. Harrison, V. Choudhury, C. Kacmar, Developing and Validating Trust Measures for e-Commerce: An Integrative Typology, Technical Report 3, 2002. URL: https://about.jstor.org/terms.
[19] H. Rahimi, H. El Bakkali, A New Reputation Algorithm for Evaluating Trustworthiness in E-Commerce Context, in: 2013 National Security Days (JNS3), Institute of Electrical and Electronics Engineers, Rabat, Morocco, 2013, pp. 1–6. doi:10.1109/JNS3.2013.6595455.
[20] A. Beldad, M. De Jong, M. Steehouder, How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust, Computers in Human Behavior 26 (2010) 857–869. doi:10.1016/j.chb.2010.03.013.
[21] D. Prinster, A. Mahmood, S. Saria, J. Jeudy, C. T. Lin, P. H. Yi, C. M. Huang, Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI, Radiology 313 (2024). doi:10.1148/radiol.233261.
[22] A. Gupta, D. Basu, R. Ghantasala, S. Qiu, U. Gadiraju, To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System, in: Proceedings of the ACM Web Conference 2022, Association for Computing Machinery, Inc, Virtual Event, Lyon, France, 2022, pp. 3531–3540. doi:10.1145/3485447.3512248.
[23] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an Empirically Determined Scale of Trust in Automated Systems, International Journal of Cognitive Ergonomics 4 (2000). doi:10.1207/s15327566ijce0401{_}04.
[24] J. Jo, H. Zhang, J. Cai, N. Goyal, AI Trust Reshaping Administrative Burdens: Understanding Trust-Burden Dynamics in LLM-Assisted Benefits Systems, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2025, pp. 1172–1183. URL: https://dl.acm.org/doi/10.1145/3715275.3732077. doi:10.1145/3715275.3732077.
[25] A. Balayn, M. Yurrita, F. Rancourt, F. Casati, U. Gadiraju, Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2025, pp. 1–20. URL: https://dl.acm.org/doi/10.1145/3706598.3713787. doi:10.1145/3706598.3713787.
[26] M. Bollaert, O. Augereau, G. Coppin, Measuring and Calibrating Trust in Artificial Intelligence, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 14536 LNCS, Springer Science and Business Media Deutschland GmbH, 2024, pp. 232–237. doi:10.1007/978-3-031-61698-3{_}22.
[27] A. Challapally, C. Pease, R. Raskar, P. Chari, The GenAI Divide: State of AI in Business 2025, Technical Report, Massachusetts Institute of Technology NANDA, 2025.

Bridging the AI Trustworthiness Gap between Functions and Norms (Position Paper)

Daan Di Scala[1,2,*], Sophie Lathouwers[1] and Michael van Bekkum[1]

[1]TNO Netherlands Organisation for Applied Scientific Research, Data Science Department
[2]Utrecht University, Department of Information and Computing Sciences

Abstract

Trustworthy Artificial Intelligence (TAI) is gaining traction due to regulations and functional benefits. While Functional TAI (FTAI) focuses on how to implement trustworthy systems, Normative TAI (NTAI) focuses on regulations that need to be enforced. However, gaps between FTAI and NTAI remain, making it difficult to assess trustworthiness of AI systems. We argue that a bridge is needed, specifically by introducing a conceptual language which can match FTAI and NTAI. Such a semantic language can assist developers as a framework to assess AI systems in terms of trustworthiness. It can also help stakeholders translate norms and regulations into concrete implementation steps for their systems. In this position paper, we describe the current state-of-the-art and identify the gap between FTAI and NTAI. We will discuss starting points for developing a semantic language and the envisioned effects of it. Finally, we provide key considerations and discuss future actions towards assessment of TAI.

Keywords

Trustworthy AI, AI Act, Functional AI Trustworthiness, Normative AI Trustworthiness, Conceptual Language

1. Introduction

Trustworthy Artificial Intelligence (TAI) is increasingly recognised as essential for both development and deployment of AI systems to create trust and confidence with stakeholders and end users. Various approaches emerge to make AI more trustworthy, each reflecting different perspectives. The functional perspective focuses on creating AI systems with sufficient technical reliability and safety [1]. From a normative and legal point of view, TAI involves adherence to rules, standards, and compliance frameworks such as ISO/IEC 42001, ISO/IEC 23894 and the EU AI Act [2, 3, 4]. Socially, the emphasis of trustworthiness lies on the broader societal impact of AI, which addresses issues such as power, ethics, privacy, inclusivity, and the systems' perceived benevolence [5, 6, 7, 8].

While many approaches have been proposed towards TAI, there is a clear lack of overlap and integration between functional and normative approaches. As the AI Act focuses on norms for high-risk systems, many existing initiatives provide risk assessment frameworks to check compliance [9, 10, 11, 12]. For example, the AI Risk Ontology (AIRO) [13] has been proposed to determine which systems are high-risk and to document related risk information. Although useful, these risk-based approaches focus on application categories of AI systems and remain largely disconnected from system design choices and concrete implementation steps.

The past has shown that it can be difficult for developers to build systems that comply with regulations such as the General Data Protection Regulation (GDPR) [14]. Research has shown that this is in part because developers have trouble relating normative requirements to technical implementations [15, 16]. Therefore, it was recommended to accompany the GDPR law with techniques to use for implementing each principle. We expect regulations on TAI such as the AI Act to face similar problems, as they do not provide accompanying techniques and guidelines for developers to use.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
daan.discala@tno.nl (D. Di Scala); sophie.lathouwers@tno.nl (S. Lathouwers); michael.vanbekkum@tno.nl (M. van Bekkum)
0000-0003-1548-6675 (D. Di Scala); 0000-0002-7544-447X (S. Lathouwers); 0009-0007-3009-254X (M. van Bekkum)

To bridge the gap between norms and regulations on one side and functional requirements on the other, we see the need for a mapping that 1) supports technical implementations in compliance with regulations and 2) translates legal standards into system requirements for developers. Without such a mapping, there is a risk that abstract or high-level trustworthiness principles remain disconnected from technical decisions during AI development and functional assessment of AI systems on TAI compliance proves elusive. Therefore, we propose that a standardised semantic framework should be developed that relates functional properties of AI systems to key trustworthiness requirements. As we believe that the envisioned semantic framework should build on existing frameworks, we explore related work that can serve as a starting point. We include both normative frameworks and ways to systematically explore system designs. We then provide an initial design and conclude with key considerations and directions for future work.

2. Towards a Bridging Semantic Framework

We now describe key starting points to consider for a semantic framework that bridges the gap between functional and normative AI requirements. For this, we first describe existing normative frameworks, then point to functional description frameworks, and finally introduce concepts that we argue should be included in the bridging language.

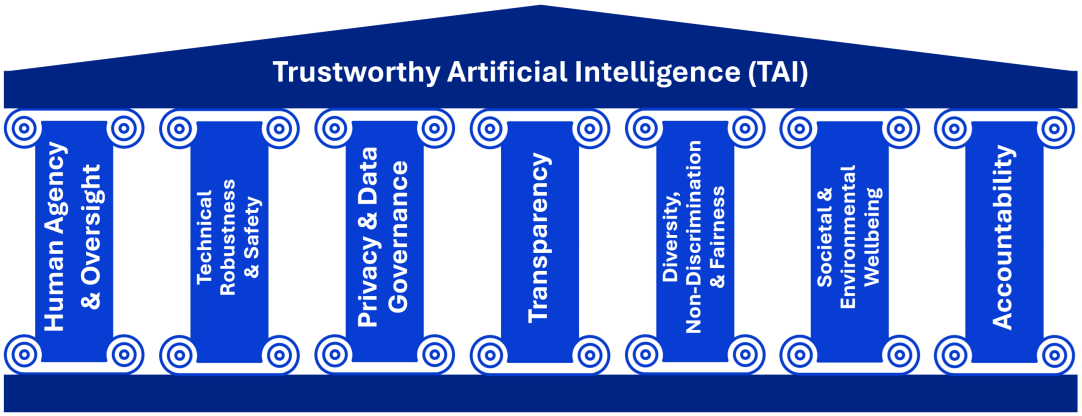

Figure 1: Seven normative key principles (pillars) of TAI as defined in [17, 18].

An important legal framework to consider is the *AI Act*, which has formally gone into effect as of August 2024 as a legal framework proposed by the EU Commission [2, 19]. The AI Act aims to promote the adoption of TAI systems by taking a risk-based approach with rules for AI developers and deployers. They identify which systems are considered high-risk AI systems. The EU HLEG has defined seven key principles or pillars for Trustworthy AI [17] as shown in Figure 1, which have been introduced in the AI Act [18]. AI systems are expected to adhere to these trustworthiness principles, under continuous evaluation throughout their life cycle. These principles are a solid starting point, as they provide definitions and goals towards TAI.

Many parties have created guidelines and assessment tools for trustworthy AI. The EU HLEG has created a list for self-assessment (ALTAI) [12], OECD introduces guidelines [20] with five values-based principles and recommendations to guide policymakers and AI actors. NIST presents the AI Risk Management Framework (AIRMF) [21]. Several tools have been specifically developed to support evaluation in terms of the AI Act [9, 10, 11, 12]. These tools can serve as inspiration, though they focus primarily on risk identification and remain high-level.

Various semantic approaches have attempted to formalise Trustworthy AI [13, 22, 23], yet they lack explicit, unambiguous definitions on terminology like Transparency. ALTAI provides a glossary for many terms related to Trustworthy AI, but these lack more rigorous semantics and still provide room for interpretation. The aforementioned AIRO does capture terms like AISystem and stakeholders

(Provider, Developer, Deployer) connected to norm sources. The Trustworthy Intelligent Systems Ontology (TISO) [22] includes terms like Safety and Transparency, but these are not connected to any normative definitions. Another approach is that of Lewis et al. [24] who choose to model based on Activities, Entities and Agents, rather than characteristics such as transparency, as these are often not well-defined.

From the other perspective, multiple functional frameworks allow us to describe AI systems. To properly assess the trustworthiness of an AI system, one needs to know the system itself. It is therefore important to explore the design of the system to identify the key components and methods that are used within the system. To ensure adoption, we recommend investigating techniques that are familiar to developers, such as UML-based modelling languages and extracting functional requirements from user stories [1, 25, 26, 27]. For system descriptions, including system behaviour, formal semantics such as the System Ontology exists [28]. For a specific focus on AI systems, system modelling approaches such as Boxology [29] are suitable because they lay the ground work for semantically defining AI components.

Closely related to what we envision is the Artificial Intelligence Trust Framework and Maturity Model (AI-TMM) [30]. This framework mentions evaluation options for trustworthiness properties. However, it does not provide a systematic approach for developers to explore their system. Moreover, trustworthiness properties are not linked to existing normative frameworks.

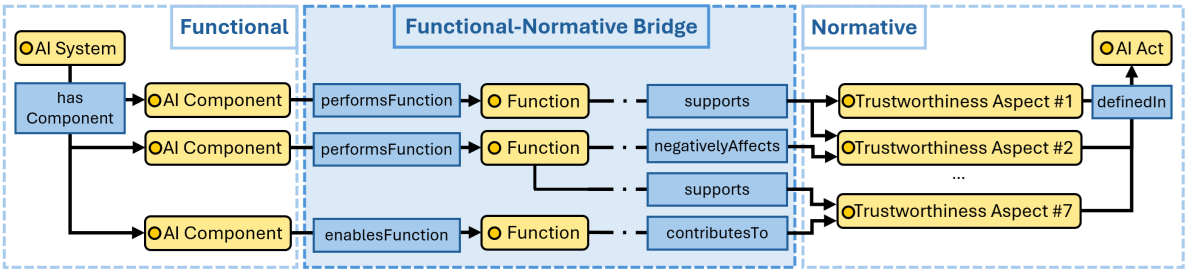

Figure 2: Conceptual language bridging functional aspects of AI components to normative (AI Act's) trustworthiness principles. Yellow blocks denote classes and blue labelled arrows denote relations.

We envision our approach for mapping from functional requirements to normative frameworks to be done in a way similar to what is described in Figure 2. This shows an approach in which each AI system's component is described in terms of functionality (performsFunction Function), which is then mapped to trustworthiness aspects (TA) from various norms such as the AI Act's seven principles. For example, a recommender system can be used as AI component to perform recommendations (performsFunction), which can support technical robustness (supports TA#1) by maintaining consistence, while it might risk being detrimental to bias (negativelyAffects TA#2) by over-relying on popular items leading to a lack of diversity. We believe that besides basing this framework on related frameworks, additional building blocks are required, threefold: 1) The language should support a broad array of functional aspects. Instances of Function could be ExtractData, InferFacts, SegmentText, SearchSemantically, SummarizeText, PredictLabels, DetectObject, RetrieveRelatedKnowledge, RankItems, or ClassifyImage. Having these functions predetermined helps developers uniformly describe AI systems. 2) As AI components do not operate in a vacuum, it is important for the language to describe the system on top of just its Functions. Describing connections between AI components and from components to data is helpful. So, it is more insightful to include descriptions such has hasInput, hasOutput, dataFormat, dataOrigin, as well as different Function relations such as evaluatedByFunction or isDependentOnFunction. 3) Ample relations should be included to tie the functional descriptions to trustworthiness aspects. Not only positive relations (like supports, contributesTo) but also negative relations (e.g., negativelyAffects) or even uncertain relations (e.g., possiblyAffects). This to properly convey the possible impact of the AI system, without having to denote it just in terms of risk or positive influence.

3. Discussion

Finally, we outline several considerations and future actions for introducing a conceptual language to describe AI systems in terms of trustworthiness. One key consideration is that this language should not aim to provide a perfect delineation or definitive classification. It should not serve as a stamp of approval nor as a guarantee of correctness as such claims would be overly strong. Instead of being positioned as a self-assessment instrument, it should be an insight tool that supports and informs assessment processes and guides system development. Another important consideration is the inherent challenge of dealing with the dynamics of the gap we have identified. It can be difficult to exhaustively identify all AI system types and to create a fully comprehensive framework of functionalities as AI systems/functionality is an area that is constantly changing. Apart from technical developments, the field of trustworthiness is also continuously growing, requiring ongoing integration of new insights. Moreover, circular dependencies may occur as normative requirements may change based on technical limitations and vice versa. To deal with the ever-changing nature of the normative and functional requirements, we propose a targeted approach by e.g. initially focusing on specific types of AI systems and by focusing on a select number of pillars of the AI Act. This can then be updated and extended based on the users' needs. This will ensure that the tool remains practical and manageable while still offering meaningful insights.

Looking ahead, we envision a scenario in which the conceptual language has been fully developed and embedded within a practical tool. Suppose a provider is tasked to describe their AI system, either as part of the development process or as a request from the deployer. Using the tool, they enter the system characteristics, after which the system automatically links to the underlying trust ontology. Based on this, the tool generates insights into the functional trustworthiness of the system in accordance with relevant standards and legislation. This helps guide the developer towards assessment of their systems' trustworthiness.

To advance towards this vision, future work needs to focus on several key directions. First, what we present here is an initial design and provides starting points from different points of view. To extend this, we expect to conduct a more comprehensive analysis of existing frameworks to identify appropriate building blocks for the conceptual language. Secondly, we intend to further develop and formalise the conceptual language in a standardised format (e.g., RDF). Finally, we would like to iteratively refine the framework based on validation and evaluation in real-world use cases.

Bridging this trustworthiness gap is crucial for ensuring that AI systems are not only functionally sound but also aligned with societal values and legal expectations. We believe that a positive impact can be made by helping stakeholders towards creating TAI according to standards, by providing understandable insights into the trustworthiness of their systems.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, J. Grundy, Requirements engineering for artificial intelligence systems: A systematic mapping study, 2022. URL: https://arxiv.org/abs/2212.10693. arXiv:2212.10693.

[2] European Union, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM/2021/206final (2021) 1–107.

[3] S. A. Benraouane, AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems, Productivity Press, 2024.

[4] A. Simonetta, M. C. Paoletti, Iso/iec standards and design of an artificial intelligence system (2024).

[5] A. Hagerty, I. Rubinov, Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence, arXiv preprint arXiv:1907.07892 (2019).
[6] J. G. O. Marko, C. D. Neagu, P. B. Anand, Examining inclusivity: the use of ai and diverse populations in health and social care: a systematic review, BMC Medical Informatics and Decision Making 25 (2025) 57. doi:10.1186/s12911-025-02884-1.
[7] B. Memarian, T. Doleck, Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review, Computers and Education: Artificial Intelligence 5 (2023) 100152. doi:10.1016/j.caeai.2023.100152.
[8] E. Novozhilova, K. Mays, S. Paik, J. E. Katz, More capable, less benevolent: Trust perceptions of AI systems across societal contexts, Machine Learning and Knowledge Extraction 6 (2024) 342–366. doi:10.3390/make6010017.
[9] Future of Life Institute, EU AI Act Compliance Checker, 2025. Available at: https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/.
[10] Stichting Algorithm Audit, AI Act Implementation Tool, 2025. Available at: https://algorithmaudit.eu/technical-tools/implementation-tool/.
[11] Ministry of Infrastructure and Water Management, AI Impact Assessment, 2024. Available at: https://www.government.nl/binaries/government/documenten/publications/2023/03/02/ai-impact-assessment/2024-IWM-AI-Impact-assessment-2.0-EN.pdf.
[12] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 2020. Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[13] D. Golpayegani, H. J. Pandit, D. Lewis, AIRO: an ontology for representing AI risks based on the proposed EU AI act and ISO risk management standards, in: A. Dimou, S. Neumaier, T. Pellegrini, S. Vahdati (Eds.), Towards a Knowledge-Aware AI - SEMANTiCS 2022 - Proceedings of the 18th International Conference on Semantic Systems, 13-15 September 2022, Vienna, Austria, volume 55 of *Studies on the Semantic Web*, IOS Press, 2022, pp. 51–65. doi:10.3233/SSW220008.
[14] S. Sirur, J. R. Nurse, H. Webb, Are we there yet? understanding the challenges faced in complying with the General Data Protection Regulation (GDPR), in: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, MPS '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 88–95. doi:10.1145/3267357.3267368.
[15] A. Senarath, N. A. G. Arachchilage, Why developers cannot embed privacy into software systems? an empirical investigation, in: Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018, EASE '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 211–216. doi:10.1145/3210459.3210484.
[16] A. Alhazmi, N. A. G. Arachchilage, I'm all ears! listening to software developers on putting gdpr principles into software development practice, Personal and Ubiquitous Computing 25 (2021) 879–892. doi:10.1007/s00779-021-01544-1.
[17] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines For Trustworthy AI, 2019.
[18] AI Act Section 3-2, 2025. Available at: https://artificialintelligenceact.eu/section/3-2/.
[19] European Commission, European legal framework for AI, 2025. Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
[20] OECD, AI principles, 2019. Available at: https://www.oecd.org/en/topics/sub-issues/ai-principles.html.
[21] National Institute of Standards and Technology (NIST), Artificial intelligence risk management framework (AI RMF 1.0), Technical Report, U.S. Department of Commerce, 2023. doi:10.6028/NIST.AI.100-1.
[22] J. I. Olszewska, Trustworthy intelligent systems: An ontological model, in: Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD, INSTICC, SciTePress, 2022, pp. 207–214. doi:10.5220/0011552700003335.
[23] J. Newman, A taxonomy of trustworthiness for artificial intelligence, CLTC: North Charleston, SC, USA 1 (2023).

[24] D. Lewis, D. Filip, H. J. Pandit, An ontology for standardising trustworthy ai, in: A. G. Hessami, P. Shaw (Eds.), Factoring Ethics in Technology, Policy Making, Regulation and AI, IntechOpen, Rijeka, 2021. URL: https://doi.org/10.5772/intechopen.97478. doi:10.5772/intechopen.97478.

[25] G. Lucassen, F. Dalpiaz, J. M. E. van der Werf, S. Brinkkemper, Improving agile requirements: the quality user story framework and tool, Requirements engineering 21 (2016) 383–403. doi:10.1007/s00766-016-0250-x.

[26] M. M. I. Molla, J. Ahmad, W. M. N. W. Kadir, A comparison of transforming the user stories and functional requirements into uml use case diagram, International Journal of Innovative Computing 14 (2024) 29–36. doi:10.11113/ijic.v14n1.463.

[27] R. Malan, D. Bredemeyer, et al., Functional requirements and use cases, Bredemeyer Consulting (2001) 335–1653.

[28] R. F. Calhau, T. P. Sales, G. Guizzardi, J. P. A. Almeida, Exploring system behavior in a system ontology, in: 42nd International Conference on Conceptual Modeling, ER 2023, CEUR, 2023.

[29] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali, A. ten Teije, Modular design patterns for hybrid learning and reasoning systems, Appl. Intell. 51 (2021) 6528–6546. URL: https://doi.org/10.1007/s10489-021-02394-3. doi:10.1007/S10489-021-02394-3.

[30] M. Mylrea, N. Robinson, Artificial intelligence (ai) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical ai, Entropy 25 (2023). URL: https://www.mdpi.com/1099-4300/25/10/1429. doi:10.3390/e25101429.

Actionable Trustworthy AI with a Knowledge-based Debugger (Position Paper)

Priyabanta Sandulu[1,*], Andrea Šipka[1,2], Sergey Redyuk[1] and Sebastian J. Vollmer[1,2]

[1]German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2]Department of Computer Science, Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Kaiserslautern, Germany

Abstract

The rapidly evolving regulatory landscape in AI presents significant challenges to establishing and maintaining trust. AI practitioners face a substantial burden in understanding and operationalizing abstract requirements. Existing solutions often lack concrete strategies for effective risk mitigation. We address these gaps by proposing an AI debugger, powered by an expandable knowledge base, that identifies risks and suggests actionable mitigation with little overhead to the end-user. A Human-in-the-Loop component supports adaptive decision-making, and the unique Requirement & Knowledge Engineering pipeline suggests the mapping between abstract guidelines and actionable specifications, pending validation by the end-user. Our framework aims to reduce the compliance overhead and streamline the development of trustworthy AI systems.

Keywords

Trustworthy AI, AI Governance, AI Risk, Risk Mitigation, Human-in-the-loop, AI Debugger

1. Practical Challenges of Trustworthy AI

Artificial Intelligence (AI) continues to redefine the boundaries of what technology can achieve. With its rapid widespread adoption, it brings both opportunities and risks. In response, the European Commission appointed a High-Level Expert Group on AI (AI HLEG) [1] to provide the ethics guideline and the assessment list for trustworthy AI (ALTAI) [2], addressing seven key requirements for trustworthy AI (tAI). These guidelines aim to direct both technical and non-technical stakeholders, and involve AI designers, developers, data scientists, procurement officers, front-end staff, legal/compliance officers, and management. Globally, many organizations proposed frameworks such as NIST [3, 4], OECD [5], the Global Partnership on Artificial Intelligence mandate [6], the General-Purpose AI Code of Practice [7], and AI Safety Institute approach [8], with overlapping or complementary goals. Despite these comprehensive frameworks, building tAI presents several interconnected challenges:

- *The Dynamic Landscape and Expertise Gap* - tAI is inherently a moving target. While foundational principles provide a strong starting point, new expectations continue to evolve across jurisdictions and industries. Each of these foundational principles covers multiple research fields, making it challenging for an individual to develop sufficient expertise across all areas simultaneously. This burden is compounded by the rapid evolution of standards (e.g., ISO/IEC 42001, 42005, 5259, 23894, 5338, 5339 [9, 10, 11, 12, 13, 14, 15, 16]), demanding continuous knowledge curation from practitioners with limited time and resources. Small and medium-sized enterprises (SMEs) face even greater challenges, as they often lack the capacity to hire domain experts or manage ongoing compliance demands [17]. While ALTAI advises seeking outside counsel, it is not always practical for SMEs to afford such dedicated support.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.
*Corresponding author: priyabanta.sandulu@dfki.de
priyabanta.sandulu@dfki.de (P. Sandulu); andrea.sipka@dfki.de (A. Šipka); sergey.redyuk@dfki.de (S. Redyuk); sebastian.vollmer@dfki.de (S. J. Vollmer)
0009-0003-9284-5093 (P. Sandulu); 0000-0002-2936-7725 (A. Šipka); 0000-0001-7131-745X (S. Redyuk); 0000-0003-2831-1401 (S. J. Vollmer)

- *Conflicting Priorities and Operationalization Challenges* - Achieving trustworthiness across all dimensions often involves inherent trade-offs, where improving one aspect might impact another. Organizations are primarily driven by the need for quick and affordable deployment. Proposing and implementing a comprehensive and resource-intensive trustworthiness initiative is often difficult to operationalize. This highlights a need for solutions that are fast, affordable, and up-to-date, maximizing automation. Where automation is not feasible, intelligent assistance and targeted education become essential. Moreover, assistance and recommendations must adapt to the specific business context and use case characteristics.

- *Lifecycle Ambiguity* - Processes like CRISP-DM [18] or KDD [19] are widely used in the context of AI system development. Yet, these models do not directly address trustworthiness or discrimination prevention [17]. While domain-specific [20, 21] or refined CRISP-DM [22] models exist, they are not universally applicable. Real-world AI systems may not strictly follow any single procedural model, and switching between frameworks is challenging [17]. Consequently, a one-size-fits-all solution tied to static system lifecycle proves insufficient. While some solutions integrate with existing life cycles [23], effective tAI solutions should be modular, independent of specific lifecycle stages, and capable of supporting hybrid, evolving workflows without tight coupling.

- *Communication and Interpretation Gap* - tAI is fundamentally a socio-technical problem. Real-world investigations are methodologically challenging due to human factors and how AI systems operate within complex socio-technical contexts [24]. Operational issues often arise from these dimensions, which are inherently more difficult to automate. Ignoring these aspects would limit our ability to build trust or effectively support stakeholders. A significant challenge stemming from this socio-technical complexity lies in communication: technical stakeholders struggle to interpret and translate legal requirements into actionable engineering specifications. This is evident from the fact that 79% of technical workers explicitly demand concrete, executable resources regarding ethical considerations [25]. Conversely, non-technical roles find it difficult to evaluate technical compliance [26]. This communication gap underscores the need for governance checks that engage stakeholders at all levels. This gap is exemplified by regulatory acts like the EU AI Act, where abstract mandates make it challenging to derive precise, unambiguous requirements for AI system design and evaluation [17]. This task makes practical implementation challenging, and could lead to avoidance. It is essential to bridge this gap by developing systematic approaches to extract, formalize, and operationalize these requirements from unstructured regulatory and ethical documentation.

- *Limited Risk Mitigation* - Another significant bottleneck in current tAI practices is the limited focus on actionable risk resolution. While state-of-the-art solutions are increasingly adept at identifying tAI risks, they often do not provide concrete, expert-guided strategies for mitigating these identified issues. Some standards, e.g., ISO/IEC 42001 on AI management systems, explicitly avoid any specific guidance on management processes and recommend combining "generally accepted frameworks, other International Standards and own experience to implement [appropriate, use-case-specific] processes such as risk management, life cycle management and data quality management". Consequently, stakeholders, particularly those without specialized expertise, struggle to translate risk assessments into actionable mitigation. This, in turn, hinders the practical deployment of tAI, underscoring the need for tools that connect risk identification to actionable mitigation.

2. Proposed Approach

We propose a Human-In-The-Loop debugger powered by an expandable knowledge base that supports the entire AI development lifecycle. This approach combines continuous human-machine collaboration with feedback loops to validate automated suggestions. It addresses two central concerns for tAI practices: **Requirement & Knowledge Engineering** to articulate trustworthiness requirements in a way that is both intuitive for human stakeholders and machine-interpretable; and **Continuous**

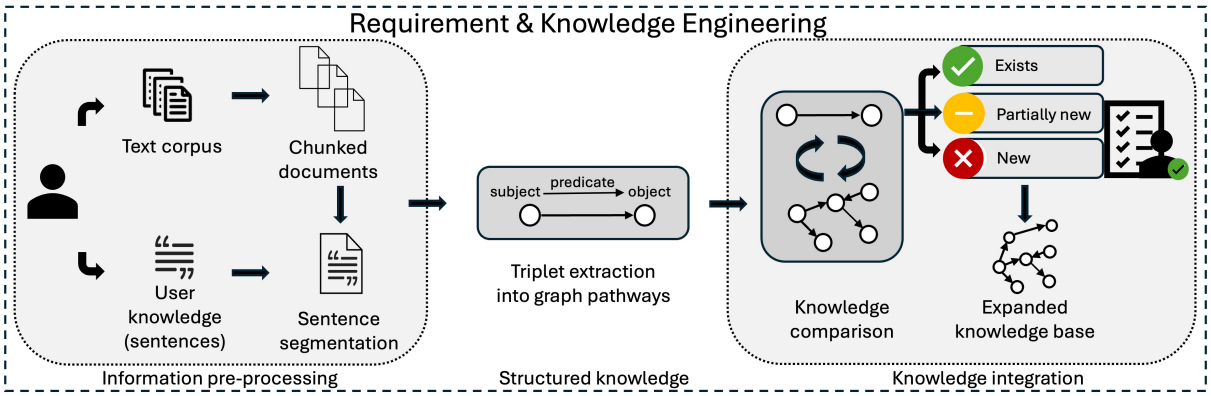

Figure 1: A flow diagram illustrating the process for extracting, structuring, and integrating tAI requirements into an expandable KB. Based on the extracted text and processed user input, the system generates new graph pathways, compares them to the existing knowledge, and uses human oversight to continuously expand KB.

Compliance to identify new risks in the AI system components and propose actionable mitigation within the entire AI development process.

2.1. Our Vision for Operationalizing Trustworthy AI

- *A Requirement & Knowledge Engineering Pipeline*: We propose a pipeline to systematically transform abstract requirements from regulatory and ethical guidelines into structured, actionable specifications. This process extracts information in the form of graph triplets and constructs a graph-based knowledge pathway, directly addressing the practical challenge of operationalizing vague mandates.
- *An AI Debugger for Actionable Risk Mitigation*: We introduce an AI debugger that goes beyond simple risk identification and provides concrete, expert-guided, and actionable mitigation strategies. Powered by an expandable knowledge base (KB), it not only identifies trustworthiness risks in AI system components but also maps them to structured mitigation pathways, and suggests context-specific remediation steps.
- *Human-in-the-Loop (HITL) Integration*: Our framework integrates a HITL component to facilitate adaptive decision-making and continuous collaboration. HITL is crucial for validating new knowledge before integration into the KB, and for approving the risks and mitigations identified by the debugger.
- *A Modular and Lifecycle-Independent Framework*: The proposed approach is designed to be modular and independent of specific AI development lifecycles. This ensures the tools can support hybrid and evolving workflows without tight coupling to a static procedural model.

2.2. Requirement & Knowledge Engineering

For AI to be trustworthy, practitioners must clearly understand and implement often abstract requirements found in guideline documents. To the best of our knowledge, state-of-the-art solutions currently lack effective methods for extracting and managing evolving tAI requirements [27, 28, 29]. We therefore propose a requirement and knowledge engineering pipeline (Figure 1) that consists of information collection and pre-processing, accepting inputs from large text corpora (like regulatory documents) and user-provided knowledge; Large language model-based segmentation of documents into manageable chunks and individual sentences [30, 31]; transform processed text segments into structured subject-predicate-object triplets [32, 33, 34], supported by co-reference resolution to handle implicit references (e.g., pronouns). The resulting triplets are normalized against a controlled schema of AI lifecycle components and form graph pathways for comparison with existing KB structures [35, 36, 37]. Unmatched triplets are considered new and validated by the end-user before integration into KB, ensuring human-in-the-loop oversight.

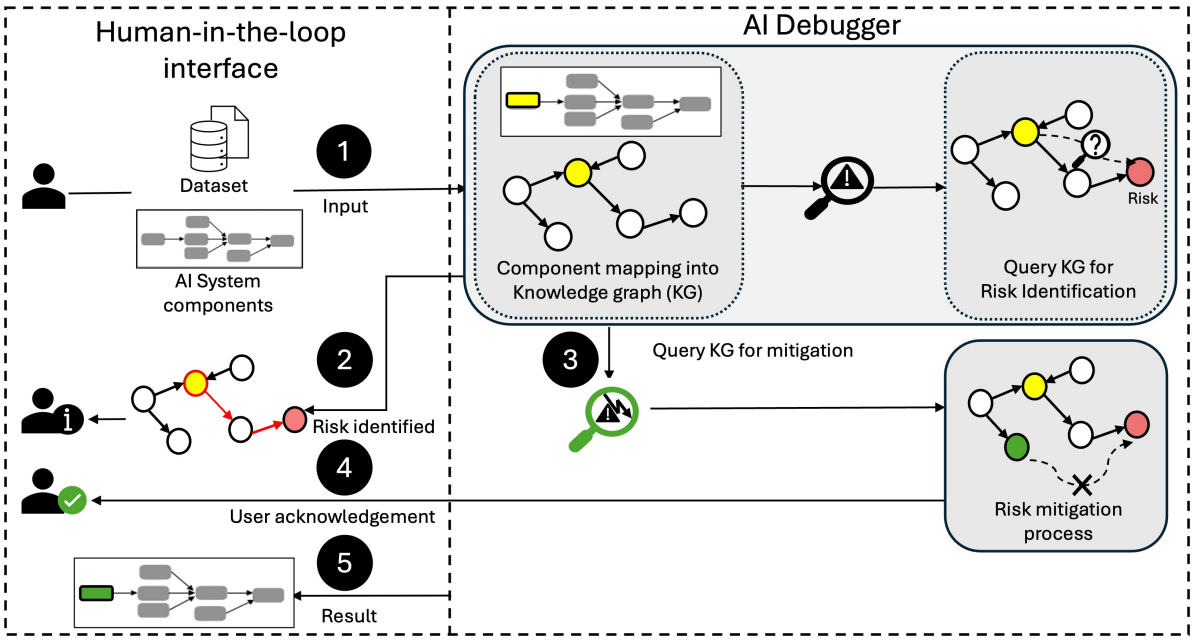

Figure 2: A flowchart depicting the **AI debugger** workflow. This process maps user inputs to the KG and query for potential risk identification and mitigation under HITL supervision.

2.3. AI Debugger

The AI debugger offers practical assistance for tAI development through an iterative workflow. It is powered by an underlying KB, a triplet-based graph repository, that contains comprehensive information on tAI dimensions (e.g., fairness, robustness), their definitions, associated metrics, identified weaknesses in models and data, and known mitigation strategies. KB is designed to be continuously updated with regulatory insights, real-world case studies, best practices, and integrates data science and AI ontologies, cross-sectoral principles, stakeholder feedback, and tool-specific compliance data. Mitigation strategies vary across domain and datasets. Accordingly, this position paper outline here on architectural level rather than a fixed catalogue.

The debugger workflow (Figure 2) begins when the end-user provides input about their dataset and AI system components. The debugger then maps these components to the structured schema of the AI toolbox and queries KB to identify known, 'potentially relevant' risks. For each identified risk, the debugger retrieves known mitigation actions from the KB contextualized by the user input. The system then initiates the risk mitigation process, which involves evaluating risk relevance under context-dependent conditions. Subsequently, the debugger informs the end-user about identified risks and proposed mitigations. Automated changes can be performed to the AI system upon explicit user confirmation; when automation is not possible, the system provides detailed feedback and steps for manual remediation. This workflow describes one iteration of the HITL interaction, which repeats until the end-user considers the system to have reached sufficient compliance.

Our proposed approach provides a roadmap to operationalize trustworthy AI, directly addressing the complexities of abstract requirements and the overhead to the end-user. We believe in the potential of these initiatives to lay the groundwork for future research and the implementation of practical compliance mechanisms for tAI systems.

Acknowledgements. This work is funded by the German Federal Ministry for Digital and Transport (BMDV) as part of the project *MISSION KI - Nationale Initiative für Künstliche Intelligenz und Datenökonomie* (45KI22B021).

Declaration on Generative AI. During the preparation of this work, the author(s) used Gemini-2.5 Flash in order to: conduct grammar and spelling check, paraphrase and reword, improve writing style. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] European Commission, High-level expert group on artificial intelligence (ai), 2020. URL: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence, accessed: 25 July 2025.

[2] European Commission, Assessment list for trustworthy artificial intelligence (altai) for self-assessment, 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, accessed: 25 July 2025.

[3] E. Tabassi, Artificial intelligence risk management framework (ai rmf 1.0) (2023).

[4] Artificial Intelligence Risk Management Framework: Generative AI Profile, Technical Report NIST AI 600-1, National Institute of Standards and Technology (NIST), 2024. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf, accessed: 10 September 2025.

[5] B. OECD, Recommendation of the council on artificial intelligence, Organisation for Economic Cooperation and Development (2019).

[6] M. Saoner, G. FRANCA, Global partnership on artificial intelligence: the future of work (2020).

[7] European AI Office, Drawing-up a general-purpose ai code of practice, 2025. URL: https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice, accessed: 25 July 2025.

[8] Department for Science, Innovation and Technology, A. U. Kingdom, Ai safety institute approach to evaluations, 2024. URL: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations, accessed: 10 September 2025.

[9] Iso/iec 42001: Artificial intelligence — management system, 2023. URL: https://www.iso.org/standard/81230.html, accessed: 10 September 2025.

[10] International Organization for Standardization, International Electrotechnical Commission, ISO/IEC 42005:2023 – Artificial Intelligence – Guidance for AI impact assessment, 2023.

[11] Iso/iec 5259-1: Artificial intelligence — data quality for analytics and machine learning — part 1: Overview, terminology and examples, 2024. URL: https://www.iso.org/standard/81088.html, accessed: 10 September 2025.

[12] Iso/iec 5259-3: Artificial intelligence — data quality for analytics and machine learning — part 3: Process requirements, 2024. URL: https://www.iso.org/standard/81092.html, accessed: 10 September 2025.

[13] Iso/iec 5259-4: Artificial intelligence — data quality for analytics and machine learning — part 4: Process framework, 2024. URL: https://www.iso.org/standard/81093.html, accessed: 10 September 2025.

[14] Iso/iec 23894: Information technology — artificial intelligence — guidance on risk management, 2023. URL: https://www.iso.org/standard/77304.html, accessed: 10 September 2025.

[15] Iso/iec 5338: Information technology — artificial intelligence — ai system life cycle processes, 2023. URL: https://www.iso.org/standard/81118.html, accessed: 10 September 2025.

[16] Iso/iec 5339: Information technology — artificial intelligence — guidance for ai applications, 2024. URL: https://www.iso.org/standard/81120.html, accessed: 10 September 2025.

[17] H. Kortum, J. Rebstadt, T. Böschen, P. Meier, O. Thomas, Towards the operationalization of trustworthy ai: integrating the eu assessment list into a procedure model for the development and operation of ai-systems, in: INFORMATIK 2022, Gesellschaft für Informatik, Bonn, 2022, pp. 283–299.

[18] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Springer-Verlag London, UK, 2000, pp. 29–39.

[19] A. Azevedo, M. F. Santos, KDD, SEMMA and CRISP-DM: a parallel overview, IADIS European Conference Data Mining (2008) 182–185.

[20] C. Silva, M. Saraee, M. Saraee, Data science in public mental health: a new analytic framework, in: 2019 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2019, pp. 1123–1128.

[21] G. Mariscal, O. Marban, C. Fernandez, A survey of data mining and knowledge discovery process

models and methodologies, The knowledge engineering review 25 (2010) 137–166.
[22] M. Haakman, L. Cruz, H. Huijgens, A. Van Deursen, Ai lifecycle models need to be revised: An exploratory study in fintech, Empirical Software Engineering 26 (2021) 95.
[23] N. Kemmerzell, A. Schreiner, H. Khalid, M. Schalk, L. Bordoli, Towards a better understanding of evaluating trustworthiness in ai systems, ACM Computing Surveys 57 (2025) 1–38.
[24] D. Kowald, S. Scher, V. Pammer-Schindler, P. Müllner, K. Waxnegger, L. Demelius, A. Fessl, M. Toller, I. G. Mendoza Estrada, I. Šimić, et al., Establishing and evaluating trustworthy ai: overview and research challenges, Frontiers in Big Data 7 (2024) 1467222.
[25] C. Miller, R. Coldicott, People, power and technology: The tech workers' view, 2019. URL: https://doteveryone.org.uk/report/workersview, accessed: 25 July 2025.
[26] U. Gasser, V. A. Almeida, A layered model for ai governance, IEEE Internet Computing 21 (2017) 58–62.
[27] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy ai: From principles to practices, ACM Computing Surveys 55 (2023) 1–46.
[28] K. Crockett, E. Colyer, L. Gerber, A. Latham, Building trustworthy ai solutions: A case for practical solutions for small businesses, IEEE Transactions on Artificial Intelligence 4 (2021) 778–791.
[29] M. T. Baldassarre, D. Gigante, M. Kalinowski, A. Ragone, Polaris: A framework to guide the development of trustworthy ai systems, in: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, 2024, pp. 200–210.
[30] A. V. Duarte, J. Marques, M. Graça, M. Freire, L. Li, A. L. Oliveira, Lumberchunker: Long-form narrative document segmentation, arXiv preprint arXiv:2406.17526 (2024).
[31] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, M. Schedl, Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation, arXiv preprint arXiv:2406.16678 (2024).
[32] Z. Chen, J. Liu, D. Yang, Y. Xiao, H. Xu, Z. Wang, R. Xie, Y. Xian, Exploiting duality in open information extraction with predicate prompt, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 125–133.
[33] A. Papaluca, D. Krefl, S. M. Rodriguez, A. Lensky, H. Suominen, Zero-and few-shots knowledge graph triplet extraction with large language models, arXiv preprint arXiv:2312.01954 (2023).
[34] C. Niklaus, M. Cetto, A. Freitas, S. Handschuh, A survey on open information extraction, arXiv preprint arXiv:1806.05599 (2018).
[35] K. Xu, L. Wang, M. Yu, Y. Feng, Y. Song, Z. Wang, D. Yu, Cross-lingual knowledge graph alignment via graph matching neural network, arXiv preprint arXiv:1905.11605 (2019).
[36] S. Hertling, J. Portisch, H. Paulheim, Kermit–a transformer-based approach for knowledge graph matching, arXiv preprint arXiv:2204.13931 (2022).
[37] H. Bunke, Graph matching: Theoretical foundations, algorithms, and applications, in: Proc. Vision Interface, volume 2000, 2000, pp. 82–88.

A Risk Index to Guide Responsible Adoption of Artificial Intelligence

Mahboubehsadat Jazayeri[1,*], Paolo Ceravolo[2] and Samira Maghool[3]

[1]Mahboubehsadat.jazayeri@unimi.it, Italy
[2]Paolo.Ceravolo@unimi.it, Italy
[3]Samira.Maghool@unimi.it, Italy

Abstract

This paper proposes a Risk Index (RI) for evaluating AI systems by integrating compliance with Trustworthy AI principles and the EU AI Act with deployment risk factors such as complexity, domain sensitivity, and scale. AI systems are assessed across their lifecycle (pre-, in-, and post-processing) using two questionnaires: a qualitative self-assessment and a 0–5 quantitative scoring answered by multiple stakeholders. Fictional case studies in hiring and healthcare show that the RI highlights vulnerabilities missed by compliance checks alone, offering a targeted and actionable framework for responsible AI adoption.

Keywords

Risk Index, Responsible AI, Compliance Assessment, Non-Compensatory Scoring, Trustworthy AI

1. Introduction

Artificial Intelligence (AI) enables machines to perform tasks requiring human intelligence, such as learning and decision-making [1]. Its use in sensitive domains—healthcare, education, criminal justice, and hiring—raises ethical, legal, and social concerns, making responsible evaluation essential. In Europe, the EU AI Act [2] introduces a risk-based approach, while the GDPR [3] emphasizes transparency, accountability, and protection of individual rights. Applying these principles in practice, however, remains challenging [4]. We introduce a **Risk Index (RI)** that combines compliance scores with deployment risk factors (determinism, failure likelihood, domain sensitivity, and market exposure). The method relies on self-assessment and stakeholder scoring, using a *non-compensatory model* where failures in one dimension (e.g., fairness) cannot be offset elsewhere. Scores are weighted by deployment context to avoid over-penalization, and a **Cost of Remediation (CoR)** model prioritizes improvements based on feasibility.

2. METHODOLOGY: Risk Assessment through Life-cycle Evaluation

The framework provides a comprehensive evaluation of AI systems throughout their lifecycle. It combines qualitative self-assessments with a quantitative (0–5) scoring system to ensure alignment with the seven core principles of trustworthy AI and the EU AI Act. Through the combination of ethical reflection and measurable compliance metrics, as well as the involvement of a diverse range of stakeholders, the framework aims to deliver balanced and practical AI governance.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.
*Corresponding author.
first.author@example.com (M. Jazayeri); second.author@example.com (P. Ceravolo); third.author@example.com (S. Maghool)
https://firstauthor.example.com (M. Jazayeri); https://secondauthor.example.com/ (P. Ceravolo); thirdauthor.example.com (S. Maghool)
0000-0000-0000-0000 (M. Jazayeri); 0000-0000-0000-0000 (P. Ceravolo); 0000-0000-0000-0000 (S. Maghool)

2.1. AI Lifecycle Structure

Evaluating all AI stages traces ethical risks and enables targeted mitigation.

- **Pre-processing** data collection and preparation, where risks of bias, exclusion, and privacy violations arise.
- **In-processing** model design and training, emphasizing fairness, robustness, and transparency.
- **Post-processing** deployment and monitoring, where accountability is critical.

While compliance scores offer a structured assessment, they do not capture real-world deployment risks. The next step is to interpret these scores in context, considering where and how the system is used [1].

3. Risk Quantification

Compliance scores indicate alignment with standards but not practical risk. We propose a Risk Index (RI) combining compliance with deployment factors, inspired by ALTAI, OECD, and ISO 31000. It aligns with the EU AI Act's risk-based categorization [5] and considers four dimensions:

- **Algorithmic Determinism (D):**Predictability of system behavior. Deterministic models are easier to audit, while neural or generative models add uncertainty. We use $1 - D$ to capture higher risk from lower determinism [6].
- **Probability of Dysfunction (P):** Likelihood of harmful outputs, proxied as $1 -$ Accuracy. High accuracy does not guarantee low risk, especially with fairness or generalization issues [7].
- **Service Impact (I):** Importance of the domain; healthcare and finance are riskier than entertainment [8].
- **Market Exposure (M):** Scale or reach of the system. Larger deployments amplify potential harms [9].

Consistent with MCDA (Multi-Criteria Decision Analysis) and software risk modeling [10], the composite Risk Index is calculated as:

$$RI = (1 - C_i) \cdot (w_D \cdot (1 - D) + w_P \cdot P + w_I \cdot I + w_M \cdot M) \tag{1}$$

Here $C_i \in [0, 1]$ is the compliance score, and w_D, w_P, w_I, w_M are sectoral weights. Aggregation is non-compensatory, so low compliance in one dimension cannot be offset by high compliance in others. At design stage, the RI offers an *ex-ante* risk estimate to guide mitigation. Over time, *ex-post* indicators such as incidents or complaints update the profile.

3.1. Estimating the Cost of Remediation

Identifying risk is only the first step in managing AI systems. We also introduce a comparative model to estimate the cost of improving compliance: the *Cost of Remediation (CoR)*. This framework draws on effort estimation models from software engineering (e.g., COCOMO) and MCDA [10], integrating risk factors that influence implementation effort.

- **Technology Maturity (T):** Availability of proven solutions.
- **Skill Availability (S):** Ease of accessing the required expertise (e.g., fairness auditing, adversarial testing).
- **Service Customization (C):** Degree of system specialization. Highly customized services are harder to modify.
- **Update Frequency (U):** Rate at which systems evolve. More frequent updates incur repeated compliance efforts.

The comparative remediation effort is computed as:

$$CoR_i = \Delta C_i \cdot (w_T \cdot T + w_S \cdot S + w_C \cdot C + w_U \cdot U) \tag{2}$$

Here $\Delta C_i = C_i^* - C_i$. Variables are in $[0, 1]$; weights adapt to project constraints. The model supports prioritization: e.g., improving fairness in a proprietary system may cost more than adding transparency in a well-documented one. It guides iterative, adaptive governance.

4. Application Scenarios

We illustrate the framework with two fictional but realistic scenarios where bias and robustness are critical:

- **AI Hiring System** — Ranks applicants from CVs, raising risks of bias and fairness in employment.
- **Healthcare Risk Predictor** — Estimates readmission likelihood; accuracy is important but equity and patient impact are critical.

These scenarios test whether the scoring system identifies domain-specific risks while aligning with trustworthy AI principles.

5. Illustrating case studies

We illustrate our method with two fictional case studies. The results quantify RI and CoR, highlighting domain-specific challenges and recurring patterns in AI deployment.

Table 1
Input parameters and Risk Index (RI) calculation for Healthcare Predictor and AI Hiring System.

Parameter	Healthcare Predictor	AI Hiring System
Compliance Score (C_i)	0.76	0.78
Determinism (D)	0.20	0.20
Probability of Dysfunction (P)	0.17	0.17
Service Impact (I)	0.90	0.70
Market Exposure (M)	0.60	0.40
Penalty Term ($1 - C_i$)	0.24	0.22
Risk Index (RI)	**0.55**	**0.42**

Weights used for both cases are: $w_D = 0.8$, $w_P = 1.0$, $w_I = 1.2$, $w_M = 0.7$.

Table 2
Input parameters and Cost of Remediation (CoR) for Healthcare Predictor and AI Hiring System

Parameter	Healthcare Predictor	AI Hiring System
Compliance Score (Ci)	0.76	0.78
Target Compliance (Ci*)	0.92	0.87
Compliance Gap (ΔC_i)	0.16	0.09
Technology Maturity (T)	0.40	0.70
Skill Availability (S)	0.50	0.80
Service Customization (C)	0.90	0.60
Update Frequency (U)	0.70	0.40
Cost of Remediation (CoR)	0.50	0.24

Weights used for both cases are: $w_T = 1.2$, $w_S = 1.5$, $w_C = 1.3$, $w_U = 1.0$

All input values are fictional. RI and CoR use different weighting schemes but fixed weights were applied for comparability.

Tables 1 and 2 present the calculation of RI and CoR for selected input values. As highlighted in recent literature, model performance alone does not determine whether an algorithm is suitable for deployment. For example, in a study on emergency department admissions, Mone et al. [11] evaluated three machine learning algorithms on real hospital data (120,600 records). The Gradient Boosted Machine (GBM) achieved the highest accuracy (80.31%), followed closely by a decision tree (80.06%) and logistic regression (79.94%). While these differences appear marginal, they do not capture the broader risks associated with deploying these models in sensitive contexts.

Even with 80% accuracy, models in sensitive domains (e.g., healthcare) can yield high RI if interpretability or monitoring is lacking. Our case study shows the healthcare predictor has higher RI despite similar accuracy, due to indirect harm and limited stakeholder control.

In Table 1, we assume the GBM is adopted as the predictor. Since GBM can be deterministic under certain conditions, we assign the D parameter a value of 0.8, leading to a corresponding risk of $1-0.8 = 0.2$. The P parameter is computed as the inverse of the model's observed accuracy. Following [11], we set it to $1 - 0.83 = 0.17$. Although D and P are identical, different contextual parameters lead to different RI values. This shows that compliance scores alone are not enough, as they do not reflect real-world risk. The RI metric addresses this issue by taking into account sensitivity, impact and exposure. Consequently, a system such as a Healthcare Predictor, which is compliant, can have a higher RI than a Hiring System. This interprets the spirit of the AI Act in a more effective way, guides stakeholders in assessing the appropriateness of deployment beyond mere legal compliance.

6. Conclusion

Algorithms are not inherently good or bad; their suitability depends on context. RI supports this assessment, aligning with the EU AI Act. For example, the Healthcare Predictor poses higher societal risk than the Hiring System despite similar compliance.

References

[1] L. Floridi, Ethics of Artificial Intelligence: An Overview, Springer, 2023.

[2] P. Ceravolo, E. Damiani, M. E. D'Amico, B. d. T. Erb, S. Favaro, N. Fiano, P. Gambatesa, S. La Porta, S. Maghool, L. Mauri, et al., Hh4ai: A methodological framework for ai human rights impact assessment under the euai act, arXiv preprint arXiv:2503.18994 (2025).

[3] Regulation (EU), General data protection regulation, Intouch 25 (2018) 1–5.

[4] S. Maghool, E. Casiraghi, P. Ceravolo, Enhancing fairness and accuracy in machine learning through similarity networks, in: International conference on cooperative information systems, Springer, 2023, pp. 3–20.

[5] Eu artificial intelligence act, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R0106, 2024. Regulation 2021/0106.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, volume 30, 2017.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociological Methods & Research 50 (2021) 3–44.

[8] E. Ntoutsi, et al., Bias in data-driven artificial intelligence systems, Journal Name (2020).

[9] C. D. B. of Latin America, OECD Public Governance Reviews The Strategic and Responsible Use of Artificial Intelligence in the Public Sector of Latin America and the Caribbean, OECD Publishing, 2022.

[10] I. O. for Standardization, ISO 31000: 2018, Risk Management-Guidelines, International Organization for Standardization, 2018.

[11] M. Mone, et al., Using data mining to predict hospital admissions from the emergency department, Journal of Health Informatics in Developing Countries 10 (2016) 91–98. URL: https://pure.qub.ac.uk/files/147123726/DComp.pdf.

Trustworthy-by-Design: Building a Generative AI Chatbot for Italian Public Administration

Chandana Sree Mala[1,2,*], Gizem Gezici[2], Sezer Kutluk[2] and Fosca Giannotti[2]

[1]University of Pisa, Pisa, Italy
[2]Scuola Normale Superiore, Pisa, Italy

Abstract

This paper presents a trustworthy-by-design framework for generative AI chatbots specifically developed for the Italian public administration, grounded in the EU's Ethics Guidelines for Trustworthy AI. Leveraging the seven key requirements defined by the High-Level Expert Group (HLEG), we assess five technical dimensions—data type, usage, model accessibility, pipeline selection, and model size—to guide the framework's development. In a real-world case study involving proprietary ICT manuals from the Italian public sector, we evaluate four alternative pipeline configurations across these dimensions and identify a Retrieval-Augmented Generation (RAG)-based architecture as the most effective solution, considering both technical and trustworthiness factors. Guided by the insights and challenges of the current use case, we implemented a preliminary retrieval component and obtained initial findings. This research bridges technical design and trustworthiness imperatives, laying the groundwork for developing trustworthy generative AI systems in public administration.

Keywords

Trustworthy AI, Public Administration, LLMs, Generative AI Chatbots, EU Trustworthy AI Guidelines, RAG

1. Introduction

On April 8, 2019, the European Union introduced the ethics guidelines which articulate a framework for achieving Trustworthy artificial intelligence (AI) based on fundamental rights. This initiative was led by the EU High-Level Expert Group on AI (HLEG) [1] and has three components, which should be met throughout the system's entire life cycle: it should be lawful, ethical, and robust. The Ethics Guidelines for Trustworthy AI (EGTAI) developed by the HLEG outline four core ethical principles that must be respected: (i) respect for human autonomy, (ii) prevention of harm, (iii) fairness, and (iv) explicability. To operationalize these ethical principles, the HLEG translated them into seven key requirements designed to guide the development of trustworthy AI systems as *human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability.*

In line with this, our approach to developing a generative AI chatbot for use in public administration is grounded in the aforementioned seven requirements of trustworthy AI. Adopting a trustworthy-by-design methodology, we first assess the technical requirements of the current use case through a trustworthiness lens before proceeding to implementation. Building on this analysis, we propose a suitable schema—rather than a finalized solution, as the approach is still under development and not yet fully validated—that aligns with both technical requirements and trustworthiness principles. We then implement the selected architecture and present preliminary results.

The structure of the paper is as follows. Section 2 introduces the current use case, highlighting its main challenges and discussing the technical and trustworthiness considerations across different

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.
c.mala@studenti.unipi.it;chandana.mala@sns.it (C. S. Mala); gizem.gezici@sns.it (G. Gezici); sezer.kutluk@sns.it (S. Kutluk); fosca.giannotti@sns.it (F. Giannotti)
0009-0004-7500-6121 (C. S. Mala); 0000-0001-9782-5751 (G. Gezici); 0000-0002-3048-5526 (S. Kutluk); 0000-0003-3099-3835 (F. Giannotti)

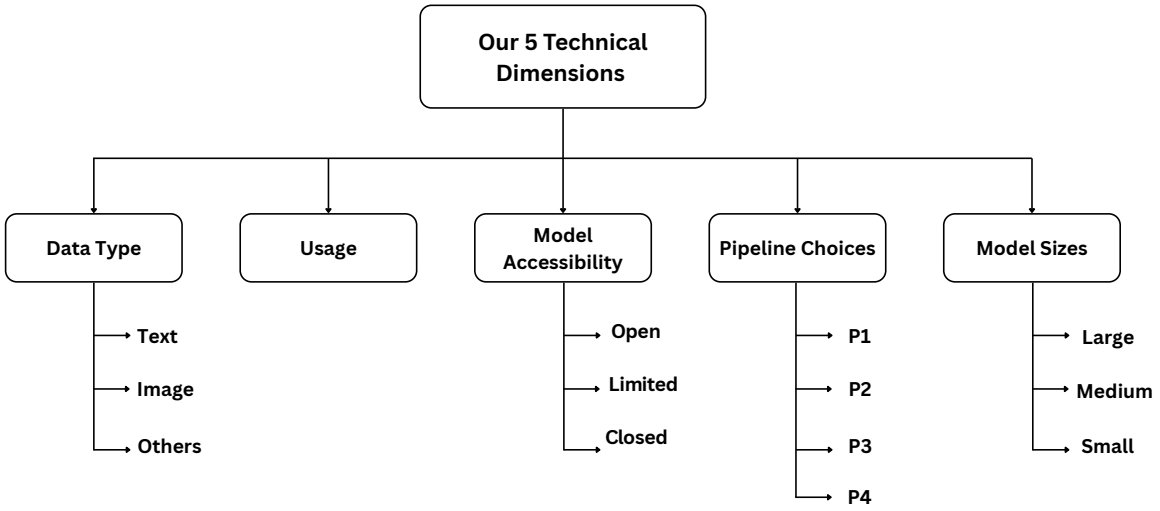

Figure 1: Our Trustworthy AI Technical Assessment Framework. The diagram illustrates our comprehensive evaluation methodology, which assesses seven trustworthy AI requirements across five key technical dimensions.

pipeline configurations. Section 3 presents the initial implementation of the proposed schema, and reports preliminary findings. Finally, Section 4 concludes the paper and outlines potential future work.

2. Methodology

We analyze the seven requirements of Trustworthy AI in relation to the design choices and technical components of our current use case. To structure this evaluation, we define five key technical dimensions—*data type, usage, model accessibility, pipeline choice*, and *model size*—as shown in Figure 1. Each trustworthy AI requirement is then assessed across these five dimensions to identify relevant challenges, trade-offs, and implementation considerations.

2.1. Current Use Case

The present use case centers on the development of a trustworthy-by-design chatbot system that leverages generative AI to address domain-specific requirements while maintaining a strong emphasis on system trustworthiness. The system targets employees of an Italian public sector government institution and operates on a proprietary dataset of technical user manuals associated with the organization's internal Information and Communication Technology (ICT) applications. Due to confidentiality, specific examples from the dataset cannot be shared. The core objective is to assist users in navigating and utilizing internal ICT systems by automatically generating helpful responses based on technical user manuals. To ensure accessibility, the chatbot must deliver answers in clear, conversational, and non-technical language, enabling individuals without specialized expertise to pose natural language queries and receive comprehensible guidance. This requires the system to effectively translate complex technical content into user-friendly responses, thereby removing the need for prior domain knowledge typically required to interpret such documentation. By improving the usability and accessibility of internal documentation, this approach aims to streamline user support and minimize reliance on direct consultation of technical manuals.

In addition to our trustworthiness analysis, we apply the AI Act's [2] risk-based assessment framework as well. Given that the use case involves a chatbot application, it is classified as a *limited risk* AI system under the AI Act. Consequently, the system must fulfill transparency requirements, including clearly informing users that they are interacting with an AI system. Furthermore, throughout its lifecycle,

the AI system must guarantee privacy and data security in accordance with regulations such as the General Data Protection Regulation (GDPR) [3]. When personal data is processed during training or deployment—which does not apply to the current use case—appropriate safeguards, such as data anonymization, should be enforced to ensure its protection.

2.2. Challenges

Ensuring the effectiveness and trustworthiness of the proposed system requires addressing several fundamental requirements specific to this use case. One of the key challenges is that the chatbot must remain aligned with evolving manuals, ensuring that its responses reflect the latest information. In addition, to build trust especially in public administration, it should provide source attributions so that users can verify the answers generated by the chatbot. At the same time, data privacy must be protected by keeping sensitive knowledge within controlled local bases rather than exposing it to external LLMs.Another important challenge lies in bridging the lexical gap between non-technical user queries and the technical terminology of the manuals—a problem widely discussed in Information Retrieval (IR). In our case, this gap stems mainly from limited domain-specific knowledge of end-users rather than from the difference in vocabulary. To address the lexical gap, techniques such as query expansion [4] (augmenting queries with domain-specific synonyms) and query rewriting [5] (paraphrasing the query using technical language) can be applied, leveraging the manuals themselves as rich sources of domain-specific terminology to improve retrieval accuracy.

2.3. Technical and Trustworthiness Considerations

Taking into account the domain-specific technical needs alongside trustworthiness considerations, we identify four possible pipeline options for building a trustworthy-by-design generative AI chatbot application in public administration domain.

Viable Pipeline Configurations. We outline four potential implementation pipelines based on the requirements of the selected use case.

- **Pipeline 1 (P1)**: Using a pre-trained large language model (LLM) via in-context learning (ICL) in few-shot settings[1].
- **Pipeline 2 (P2)**: Fine-tuning a pre-trained LLM and further use it through ICL in zero-shot[2] or few-shot settings.
- **Pipeline 3 (P3)**: P1 augmented with Retrieval-Augmented Generation (RAG), and then employed through ICL, preferably in a few-shot configuration.
- **Pipeline 4 (P4)**: P2, augmented by RAG, is further utilized through ICL in either zero-shot or few-shot settings.

Pipeline Assessment. P1 is deemed unsuitable, as the chatbot must be tailored to the specific context of public administration rather than acting as a general-purpose agent. P2 addresses this by fine-tuning a pre-trained LLM on domain-specific data, enabling adaptation to the requirements of the target application. The resulting model can then be applied in zero-shot or few-shot prompting settings via ICL to enhance task performance. Crucially, all models considered here are chat based LLMs that is, pre-trained base models (e.g., GPT-3 [6]) subsequently fine-tuned for interactive dialogue using supervised instruction datasets and reinforcement learning from human feedback (RLHF) with ChatGPT[3] being a representative example of such conversationally specialized variants.

Although pipeline **P2** offers a viable strategy, it faces notable limitations. The fine-tuning of proprietary LLMs on sensitive datasets entails significant privacy risks, and the selection among closed-source,

[1]For guiding LLMs by including a small number of task-specific examples within the prompt, enabling the model to better interpret and respond to the desired task.
[2]The model is prompted without showing any examples.
[3]https://openai.com/index/chatgpt/

open-source, and intermediate open-access models is informed not only by performance considerations but also by trustworthiness requirements. While closed-source models continue to offer state-of-the-art performance and greater ease of use, the performance gap with open models has diminished to roughly one year. [7]. As a result, factors such as suitability for the intended application, cost effectiveness, and organizational needs increasingly drive this decision, with trust, privacy, and transparency playing a decisive role in our context. Furthermore, adapting chat models through domain-specific fine-tuning requires transforming data into instruction-style input–output pairs, a labor intensive and computationally expensive process that scales with model size and dataset volume.

Based on our analysis of the current use case and associated trustworthiness considerations, we argue that a generative AI chatbot application incorporating RAG [8] represents the most suitable solution. RAG [9, 10] is a framework designed to enhance the quality of responses generated by LLMs by grounding them in external knowledge sources. It blends the encyclopedic memory of a search engine with generative modeling through two key modules: retrieval and generation. Unlike traditional models that depend solely on inherent knowledge, RAG incorporates external documents from sources such as databases, search engines, or vector stores, enabling responses that are more reliable and verifiable. Typical RAG systems [11, 12] adopt a retrieve-then-read design, where a retriever [13] identifies relevant documents and a generator conditions its response on both the query and retrieved content.

By grounding answers in external sources, RAG mitigates key limitations of standalone LLMs: hallucinations [14, 15, 16, 17], high fine-tuning costs [18], restricted context windows [19], and inherent knowledge cut-off dates [8]. Taking these factors into account, we identify **P3** as the most suitable pipeline for our use case, as it incorporates RAG while avoiding the practical limitations of fine-tuning. Nevertheless, **P4**—which combines domain-specific fine-tuning with RAG—could offer additional performance gains in contexts with abundant training data and sufficient computational resources, albeit with higher cost, complexity and data-privacy concerns.

Proposed Approach In line with the technical and trustworthiness considerations, we adopt a RAG-based pipeline (**P3**) as the foundation of our solution for building a trustworthy-by-design generative AI chatbot in the public administration domain. RAG augments the generative process of LLMs with external, verifiable knowledge [8, 9, 10], thereby mitigating limitations of standalone LLMs. Our core contribution lies in tailoring this architecture to the needs of a multilingual domain-specific setting. In particular, to tackle with the key challenges discussed in Section 2.2, the proposed pipeline is designed to:

1. **Continuously integrate updated domain manuals** without re-training, simply by refreshing the knowledge base,
2. **Enhance trust by fostering transparency** through linking retrieved sources to generated responses, and
3. **Address cross-lingual and domain-specific retrieval challenges**, especially with respect to Italian documentation, through the deployment of optimized embedding and retrieval strategies.

From a technical standpoint, the pipeline employs the standard two-step architecture: during the retrieval phase, the user query is augmented with semantically relevant passages extracted from the knowledge base, while in the generation phase, the LLM produces a response conditioned on both the query and the retrieved context. In contrast to more resource-intensive fine-tuning approaches, this configuration provides adaptability while maintaining operational efficiency.

In the following section, we systematically evaluate retrieval strategies for Italian ICT manuals—including English translation, Italian-only embeddings, and multilingual embedding models—and demonstrate how these choices affect the effectiveness of retrieval within our proposed RAG-based chatbot pipeline.

Model	Original (IT)	Translated (EN)
BERTino	0.81	–
Gattina	0.70	–
Mmarco	0.66	–
GTE-en	–	0.88
BGE	–	0.79
mGTE	0.88	**0.89**
KaLM-E	0.86	0.72

Table 1
Comparison of Retrieval Accuracy

3. Experimental Setup

To operationalize the proposed approach, we develop an initial experimental setup focused on optimizing retrieval performance, a factor that is critical for enhancing downstream generation quality in RAG pipelines [20]. The effectiveness of RAG architectures relies on both the accuracy of embedding models in representing document chunks within a high-dimensional semantic space and the capacity of the LLM to effectively leverage this contextual information during response generation. In this study, we systematically examine three strategies for generating embeddings of Italian ICT manuals for retrieval tasks: (i) translating the Italian text into English and subsequently using English-only embedding models, as outlined in [21], while maintaining connections to the original Italian sources; (ii) utilizing embedding models trained exclusively on Italian corpora; and (iii) directly applying multilingual embedding models to the Italian text. For evaluation, we rely on a $1k$ subset of a synthetic QA dataset provided by ReDix Informatica [22], which contains question–context–answer triples automatically generated from Wikipedia passages. This setup enables us to isolate retrieval behavior by focusing on question–context pairs. We compare embedding models across three categories: i) Italian-specific (Bertino [23], Gattina [24], Mmarco [25]), ii) English-only (GTE-EN [26], BGE [27]), and iii) multilingual (mGTE [28], KaLM-E [29]).

Our preliminary findings, summarized in Table 1, indicate that multilingual GTE exhibits strong cross-lingual capabilities, achieving the highest performance on both translated and original Italian text. Notably, English-only GTE-EN also outperforms Italian-trained models, which consistently show lower accuracy. These results highlight the potential of multilingual and translation-based strategies in non-English retrieval scenarios, while also underscoring the limitations of current Italian-specific embedding models. We further validated these comparative trends using our confidential Italian ICT manuals from a real-world public administration use case, confirming the robustness of the experimental setup.

4. Conclusion & Future Work

This paper introduces a trustworthy-by-design framework for building generative AI chatbots adapted to the Italian public administration, aligned with the EU's Ethics Guidelines for Trustworthy AI. It evaluates four pipeline configurations across five technical aspects, concluding that RAG offers the best balance of transparency, lowered hallucination risk, and cost efficiency without the need for constant fine-tuning. A practical case study on proprietary ICT manuals demonstrates the approach's real-world applicability. The work concentrates on the retrieval phase of RAG, highlighting its vital role in system performance, while deferring the evaluation of generation phase as future work.

Future research will expand evaluation efforts to wider contexts, including developing query expansion methods, comparing a variety of retrieval models, exploring advanced prompting techniques, and validating results across additional European languages and sectors. These efforts aim to better understand how retrieval affects generation quality and to enhance the practical utility of the framework for trustworthy, regulation-compliant AI applications in the European public sector.

References

[1] European Commission - High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2020. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, accessed: 2025-08-04.

[2] European Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, Official Proposal, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2021%3A206%3AFIN, cOM(2021) 206 final.

[3] European Parliament and Council of the European Union, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation), Official Journal of the European Union, L119, 1–88, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] S. Kuzi, A. Shtok, O. Kurland, Query expansion using word embeddings, in: Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 1929–1932.

[5] X. Ma, Y. Gong, P. He, H. Zhao, N. Duan, Query rewriting in retrieval-augmented large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5303–5315.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[7] B. Cottier, J. You, N. Martemianova, D. Owen, How far behind are open models?, 2024. URL: https://epoch.ai/blog/open-models-report, accessed: 2025-07-18.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[9] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).

[10] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meeting llms: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, 2024, pp. 6491–6501.

[11] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented language model pre-training, in: International conference on machine learning, PMLR, 2020, pp. 3929–3938.

[12] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3784–3803.

[13] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering., in: EMNLP (1), 2020, pp. 6769–6781.

[14] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55.

[15] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's song in the ai ocean: A survey on hallucination in large language models, 2023, URL https://arxiv.org/abs/2309.01219 (2024).

[16] O. Ayala, P. Bechard, Reducing hallucination in structured outputs via retrieval-augmented generation, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), 2024, pp. 228–238.

[17] C. S. Mala, G. Gezici, F. Giannotti, Hybrid retrieval for hallucination mitigation in large language models: A comparative analysis, arXiv preprint arXiv:2504.05324 (2025).

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[20] F. Tian, D. Ganguly, C. Macdonald, Is relevance propagated from retriever to generator in rag?, in: European Conference on Information Retrieval, Springer, 2025, pp. 32–48.

[21] C. Iscan, M. F. Ozara, A. Akbulut, Enhancing rag pipeline performance with translation-based embedding strategies for non-english documents, in: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2024, pp. 1–6.

[22] R. L. R. Informatica, wikipediaqa-ita: An open dataset of italian qa from wikipedia documents, https://https://huggingface.co/datasets/ReDiX/wikipediaQA-ita, 2024.

[23] @efederici, Sentence-bertino, 2022. URL: https://huggingface.co/efederici/sentence-BERTino.

[24] @mrinaldi, Flash-it-ha-classifier-cossim, 2024. URL: https://huggingface.co/mrinaldi/gattina-ha-classifier-cossim.

[25] @nickprock, Mmarco-bert-base-italian-uncased, 2023. URL: https://huggingface.co/nickprock/mmarco-bert-base-italian-uncased.

[26] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards general text embeddings with multi-stage contrastive learning, arXiv preprint arXiv:2308.03281 (2023).

[27] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-pack: Packed resources for general chinese embeddings, in: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, 2024, pp. 641–649.

[28] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, et al., mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, arXiv preprint arXiv:2407.19669 (2024).

[29] X. Hu, Z. Shan, X. Zhao, Z. Sun, Z. Liu, D. Li, S. Ye, X. Wei, Q. Chen, B. Hu, et al., Kalm-embedding: Superior training data brings a stronger embedding model, arXiv e-prints (2025) arXiv–2501.

Labelling the Trustworthiness of Medical AI

María Villalobos-Quesada*

National eHealth Living Lab (NeLL), Primary Care and Public Health Department, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA Leiden, the Netherlands.

Abstract

Artificial Intelligence (AI), particularly in healthcare, is promising, but poses risks. The potential benefits have motivated the EU to set a strategy that aims to harness AI's full potential while safeguarding fundamental rights. However, regulatory and governance frameworks have lagged behind the rapid evolution of AI. Although the EU regulation is developing, it is often seen as insufficient or inappropriate. The misalignment between the rapid medical AI innovation and the development of mechanisms to protect fundamental rights and align with societal values, introduces serious risks. For this reason, the concept of responsible or trustworthy AI is increasingly relevant. Yet, across the AI field, one consistent insight emerges: the value of trustworthy AI remains unclear. Within a market-oriented innovation ecosystem, aligning AI trustworthiness goals with market incentives may be transformative. This paper presents arguments to justify such an alignment by adopting an EU-wide *Trustworthy AI Label*. First, this *Trustworthy AI Label* could act as a means to give value to trustworthy AI. Second, it could complement the existing hard-law instruments. Third, it could contribute to improving the availability of evidence of medical AI, which is necessary to determine the quality of AI systems. The credibility and effectiveness of a *Trustworthy AI Label* will depend on robust and independent assessment mechanisms, which include the needs of diverse stakeholders, and which strike a balance between standardisation and context-specificity. Ultimately, an EU-endorsed label could serve as both a transparency and accountability tool and a market incentive, translating societal values into innovation pathways, ensuring AI contributes to the public good.

Keywords
Medical AI, trustworthy AI, responsible AI, label, AI Act.

1. Introduction

Artificial intelligence (AI) is regarded as a powerful and disruptive technology, generating signifi-cant expectations due to its transformative socio-economic potential [1]. As AI shifted beyond aca-demic and purely experimental settings, into broader societal and real-life contexts, policymakers embraced it as a means to increase efficiency, productivity, and competitiveness [2]. Within the European Union (EU), initiatives such as the Digital Single Market Strategy and the EU Digital Strategy explicitly promote AI development, framing it as an economic enabler that must be gov-erned to "achieve its full potential" [3].
Due to AI's potential positive and negative effects [2], the EU has adopted an ambitious dual-goal strategy: safeguarding citizens and their fundamental rights, while capitalising on AI's economic potential and positioning the EU as a global AI competitor [4]. This dual approach is embodied in legal frameworks applicable to AI, such as the General Data Protection Regulation (GDPR) and the Medical Devices Regulation (MDR), as well as specific AI instruments like the AI Act. The EU's AI regulatory environment continues to expand, with forthcoming instruments such as the AI Liabil-ity Directive and the European Health Data Space, which are designed to complete the EU govern-ance architecture on AI.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

* Corresponding author.
m.j.villalobos_quesada@lumc.nl; https://nell.eu/member/show/mar-a-villalobos-2
0000-0003-4930-1982

2. Medical AI

Among AI applications, medical AI is particularly significant [5]. It directly pertains to the provision of care, recognised as a fundamental right under the EU Charter of Fundamental Rights. It has an impact on health, which is recognised as a Human Right, and as a prerequisite for realising other Human Rights and well-being [6-8]. Additionally, medical AI is intended to be deployed in the service of society (AI Act), which is diverse, and where groups may prioritize and interpret values differently. The interplay between societal diversity and shared social values is a cornerstone of the EU (Treaty on European Union).

Globally, regulatory and governance frameworks have lagged behind the rapid evolution of AI, including in the healthcare sector [9]. In the EU, considerable efforts have been made to strengthen the regulatory apparatus, yet the evolving landscape seems insufficient by some or inappropriate by others [10-11]. Some sectors consider that the EU is over-regulating, and in doing so, losing the innovation race [12]. While regulation can be a mechanism to prevent harm in the strict sense (e.g., product safety regulation in the EU), it is less equipped to promote social values, as we have previously shown in the case of the AI Act [13]. In other words, regulation often serves as a threshold of minimum compliance. Additionally, some AI applications for health may not fall under the most strictly regulated categories of AI, creating legal gaps with potential safety risks [13].

The misalignment between the rapid medical AI innovation and the development of mechanisms to protect fundamental rights and AI align with societal values, introduces serious risks. For instance, AI diagnostic tools may exhibit significant bias [14], undermining EU fundamental rights, such as dignity, justice, and equality. Other risks associated with medical AI include privacy and data security concerns, lack of explainability and transparency, deskilling, job displacement, and challenges related to responsibility and liability [15]. The risks posed by AI are particularly critical for (historically) disadvantaged and vulnerable groups [2]. Ensuring that medical AI adheres to societal values and fundamental rights is therefore crucial for its positive impact on the EU landscape, and beyond.

In this scenario, the concept of *responsible AI* becomes increasingly relevant. Several organisations, including OECD, IEEE, and UNESCO, have produced guidelines and frameworks for responsible AI [16-18]. At the EU level, this concept corresponds to the notion of *trustworthy AI*, and it is grounded in the EU's Fundamental Rights. Trustworthy AI requires systems to be lawful, ethical, and robust throughout their lifecycle [2]. However, implementing trustworthiness in practice remains challenging, even with efforts to operationalise it through assessment frameworks [19]. Moreover, the phenomenon of "ethics washing", the use of ethical rhetoric without substantive practice, has become increasingly problematic [20]. Even when organisations claim to have adopted the concept of *trustworthy AI*, the absence of transparent and rigorous mechanisms, such as independent assessments or endorsements, may generate false expectations.

3. The Value Gap

Based on experience as an academic in the Netherlands, collaborating extensively with medical AI innovators across academic, public and private sectors, one consistent insight emerges: the value of ethical or *trustworthy AI* remains unclear. This affects end users, healthcare providers, regulators, and innovators alike. For small and medium-sized enterprises (SMEs), which often navigate innovation with limited resources, the absence of clear incentives makes it challenging to invest systematically in trustworthy-by-design practices and evaluations. In larger organisations, trustworthy AI is often integrated within or alongside compliance departments. While this integration can foster synergy, it risks reducing ethics to mere checklist compliance, or turning *trustworthy AI* into a branding exercise rather than a substantive commitment.

Within a market-oriented innovation ecosystem, aligning ethical goals with market incentives may be transformative. This approach must be pursued cautiously, to ensure that ethics and trustworthy AI are not turned into market goods, but as tools to prevent or mitigate risks and promote fundamental rights and values. Although this topic warrants a more in-depth discussion, this paper focuses on presenting the arguments that could justify such an alignment. I draw parallels with the EU energy label, and my first-hand experience in labelling the quality of health apps to illustrate the potential of this approach.

4. A Possible Path Forward: A *Trustworthy AI Label*

One promising strategy, that could make the concept of trustworthy AI visible to all stakeholders is the introduction of an EU-endorsed *Trustworthy AI Label*, inspired by the well-known EU Energy Label. Through mainly the Ecodesign Directive and the Regulation (EU) 2017/1369, the EU establishes obligations to label products such as household appliances, consumer electronics, and other energy-related goods [21]. Labelling has also had an impact on automobiles and buildings, for example. We have seen that the EU Energy Label has influenced both consumer behaviour and market dynamics. By providing easily interpretable information about energy efficiency, it compelled manufacturers to improve product designs, empowered consumers to make more sustainable choices, and significantly contributed to reductions in energy consumption while promoting greener market innovation across the EU. It is also considered to have had a global impact [22].

Inspired by the positive effects of the EU energy label and the Label2Enable project, which took steps toward labelling the quality of health apps based on CEN-ISO/TS 82304-2CEN-ISO/TS 82304-2 [23], I hypothesise that labelling the trustworthiness of AI systems using a standardised assessment framework could positively impact the medical AI industry and society in general.

This hypothesis is based on three main arguments. First, the value gap described above. In short, a *Trustworthy AI label* could serve as an incentive for AI manufacturers (using MDR terminology) to invest in trustworthy-by-design practices, and to document these efforts. Simultaneously, aspects of AI trustworthiness could be made visible to consumers of AI products, raising awareness and providing clear, actionable information to guide purchasing and usage decisions. If consumers learn to value and differentiate AI products based on their trustworthiness, they could create a market advantage for manufacturers investing in trustworthiness aspects, thereby nudging the AI market in that direction. If this seems unlikely, the EU Energy Label, is an interesting precedent, as it has successfully addressed a similar value gap regarding energy consumption [24].

Second, the need to complement the existing hard law instruments. As we argued above, legal frameworks only partially protect values deemed relevant in our society, and particularly their positive interpretations, such as health, well-being, and solidarity. A *Trustworthy AI Label* can be context-dependent, taking into account the local values and needs of local communities. It can provide a ground for analysing the potential effects on users who may be influenced or affected by AI systems. In the context of labelling the quality of health apps, we have demonstrated that such flexibility can be incorporated into assessment frameworks [25].

Third, the "crisis of availability of evidence" in the medical AI field. In a scoping review, we found that potential users cannot access relevant information to determine the quality of medical AI systems, either because the evidence is not made available or because it does not exist. Counterintuitively, this problem is more pronounced for CE- and FDA-approved AI systems [26]. When evidence exists, but is considered proprietary or sensitive, a *Trustworthy AI label* could also prove to be advantageous. If the assessment is conducted by a third party, the evidence does not necessarily have to be made publicly available or unrestricted. This can be a compromise between intellectual property and proprietary information, and transparency and explainability.

In this scenario, the credibility and effectiveness of a *Trustworthy AI Label* depend on robust and independent assessment mechanisms. Labelling must be grounded in evidence-based evaluation rather than statements without verifiable information. Experience with labelling the quality of health apps (based on the ISO/TS 82304-2:2021) vividly illustrates both the potential of labelling and the need for transparent, credible assessment frameworks [27]. Equally critical is that the label addresses the concerns and needs of diverse stakeholders. This point is complex, especially aiming at an EU-wide label, but we have shown that it is possible to strike a balance between standardisation and context-specificity [25]. Involving all relevant stakeholders, including end users, as well as academic, public and private actors, in co-developing a *Trustworthy AI Label* and its assessment framework, is vital to translate trustworthiness into actionable requirements and measurable outcomes. Tentatively, and because of their strong presence in the EU, the seven trustworthy AI requirements proposed by the EU High-Level Expert Group could be used as a starting point [2], but their adaptation to the AI system and context is necessary [13].

5. Conclusion

Ultimately, an EU-endorsed *Trustworthy AI Label* would serve as both a transparency and accountability tool and a market incentive, facilitating the conversion of societal values into innovation pathways. By signalling how humans have designed AI to be legally compliant, ethical and robust, in a transparent and independently verifiable manner, such a label would confer a competitive advantage to innovators who invest in trustworthiness. For end users, it would provide clarity and confidence when selecting AI-enabled medical tools. It offers a distinct opportunity to incorporate principles with design and market structures, ensuring AI contributes to the public good.

Acknowledgements

This research was conducted as part of the ZonMW-funded project 'DECIDE-VerA' (grant no. 08540122120004). My thanks to Petra Hoogendoorn, lead expert CEN-ISO/TS 82304-2 and coordinator of the Horizon Europe project Label2Enable. To André Krom, for the insightful discussions and insights regarding AI ethics.

Declaration on Generative AI

The author employed Grammarly to check spelling.

References

[1] Polak P, Anshari M. Exploring the multifaceted impacts of artificial intelligence on public organizations, business, and society. Humanit Soc Sci Commun. 2024 Oct 15;11(1):1373.

[2] AI HLEG. Ethics Guidelines for Trustworthy AI. High-Level Expert Group on AI. European Commission; 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, accessed: 2025-10-10

[3] European Commission. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS on enabling the digital transformation of health and care in the Digital Single Market; empowering citizens and building a healthier society. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52018DC0233

[4] von der Leyen. Speech by President von der Leyen at the Artificial Intelligence Action Summit. February 2025. URL: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/speech_25_471/SPEECH_25_471_EN.pdf, accessed: 2025-10-10

[5] Cohen IG, Evgeniou T, Gerke S, Minssen T. The European artificial intelligence strategy: implications and challenges for digital health. The Lancet Digital Health. 2020 July;2(7):e376–9.

[6] United Nations Committee on Economic, Social and Cultural Rights (CESCR). General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12 of the Covenant), 11 August 2000, E/C.12/2000/4). URL: https://www.ohchr.org/sites/default/files/Documents/Issues/Women/WRGS/Health/GC14.pdf, accessed: 2025-10-10

[7] Nampewo Z, Mike JH, Wolff J. Respecting, protecting and fulfilling the human right to health. Int J Equity Health. 2022 Dec;21(1):36.

[8] Ely Yamin A, Bottini Filho L, Gianella Malca C. Analysing governments' progress on the right to health. Bull World Health Organ. 2024 May 1;102(5):307–13.

[9] Busch F, Geis R, Wang YC, Kather JN, Khori NA, Makowski MR, et al. AI regulation in healthcare around the world: what is the status quo? 2025. URL: http://medrxiv.org/lookup/doi/10.1101/2025.01.25.25321061, accessed: 2025-10-10

[10] Kusche I. Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. Journal of Risk Research. 2024 May 11;1–14.

[11] Watcher S. Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond. Yale Journal of Law & Technology. 2024;26(3):671–718.
[12] Bradford A. The False Choice Between Digital Regulation and Innovation. SSRN Journal. 2024; URL: https://www.ssrn.com/abstract=4753107, accessed: 2025-10-10
[13] De Graaf T, Krom A, Colombo S, van de Pavert M, Harbers M, van Staalduinen J, et al. Conformity between values and EU legal requirements regarding an AI Clinical Decision Support System (AI-CDSS) for improved cardiovascular risk management. SRNN. 2025. URL: http://dx.doi.org/10.2139/ssrn.5239888, accessed: 2025-10-10
[14] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. Kalla M, editor. PLOS Digit Health. 2023 June 22;2(6):e0000278.
[15] Anderson B, Sutherland E. Collective Action for Responsible AI in Health, OECD. 2024. (OECD Artificial Intelligence Papers January 2024 No. 10). URL: https://www.oecd.org/en/publications/collective-action-for-responsible-ai-in-health_f2050177-en.html, accessed: 2025-10-10
[16] OECD. OECD AI Principles overview (2019). URL: https://oecd.ai/en/ai-principles, accessed: 2025-10-10
[17] Shahriari K, Shahriari M. IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC). Toronto, ON, Canada: IEEE; 2017. p. 197–201. URL: http://ieeexplore.ieee.org/document/8058187/, accessed: 2025-10-10
[18] UNESCO. Recommendation on the Ethics of Artificial Intelligence. France; 2021. Report No.: SHS/BIO/PI/2021/1. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137
[19] AI HLEG. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission. LU; 2020. URL: https://data.europa.eu/doi/10.2759/791819, accessed: 2025-10-10
[20] Wagner B. Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In: Bayamlioglu E, Baraliuc I, Janssens LAW, Hildebrandt M, editors. *BEING PROFILED: 10 Years of Profiling the European Citizen*, edited by Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens and Mireille Hildebrandt, Amsterdam: Amsterdam University Press, 2019, pp. 84-89. https://doi.org/10.1515/9789048550180-016, accessed: 2025-10-10
[21] BEING PROFILED. Amsterdam University Press; 2019. p. 84–9. URL: https://www.degruyter.com/document/doi/10.1515/9789048550180-016/html, accessed: 2025-10-10
[22] European Commission. Energy Efficient Products. n.d. [cited 20 Sept 2025] URL: https://energy-efficient-products.ec.europa.eu/index_en, accessed: 2025-10-10
[23] European Commission. Understanding the Energy Label. 2024. URL: https://energy-efficient-products.ec.europa.eu/ecodesign-and-energy-label/understanding-energy-label_en, accessed: 2025-10-10
[24] Label2Enable. Promoting a quality label for health apps! 2005. URL: https://label2enable.eu/
[25] Kesselring A. Energy labels in the European Union: Consumer inattention and producer responses. Energy Economics. 2025 Apr;144:108275.
[26] Llebot Casajuana B, Hoogendoorn P, Villalobos-Quesada M, Pratdepàdua Bufill C. Integrating CEN ISO/TS 82304-2 in the Catalan Health App Assessment Framework: Comparative Case Study. JMIR Mhealth Uhealth. 2025 June 4;13:e67858–e67858.
[27] Rakers MM, Van Buchem MM, Kucenko S, De Hond A, Kant I, Van Smeden M, et al. Availability of Evidence for Predictive Machine Learning Algorithms in Primary Care: A Systematic Review. JAMA Netw Open. 2024 Sept 12;7(9):e2432990.
[28] Hoogendoorn P, Versluis A, Van Kampen S, McCay C, Leahy M, Bijlsma M, et al. What Makes a Quality Health App—Developing a Global Research-Based Health App Quality Assessment Framework for CEN-ISO/TS 82304-2: Delphi Study. JMIR Form Res. 2023 Jan 23;7:e43905.

Legal References

Consolidated version of the **Treaty on European Union** [2016] OJ C 202/13.

Charter of Fundamental Rights of the European Union (2012/C326/02).

Directive 2009/125/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for the setting of ecodesign requirements for energy-related products (**Ecodesign Directive**) [2009] OJ L285/10.

Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (**AI Liability Directive**) COM (2022) 496 final.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (**General Data Protection Regulation, GDPR**) [2016] OJ L119/1.

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices (**Medical Devices Regulation, MDR**) [2017] OJ L117/1.

Regulation (EU) 2017/1369 of the European Parliament and of the Council of 4 July 2017 setting a framework for energy labelling and repealing Directive 2010/30/EU [2017] (**Energy Labelling Regulation**) OJ L198/1.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (**Artificial Intelligence Act**) [2024] OJ L1689/1.

Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847 [2025] (**European Health Data Space Regulation**) OJ L (327) 5 March 2025.