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**

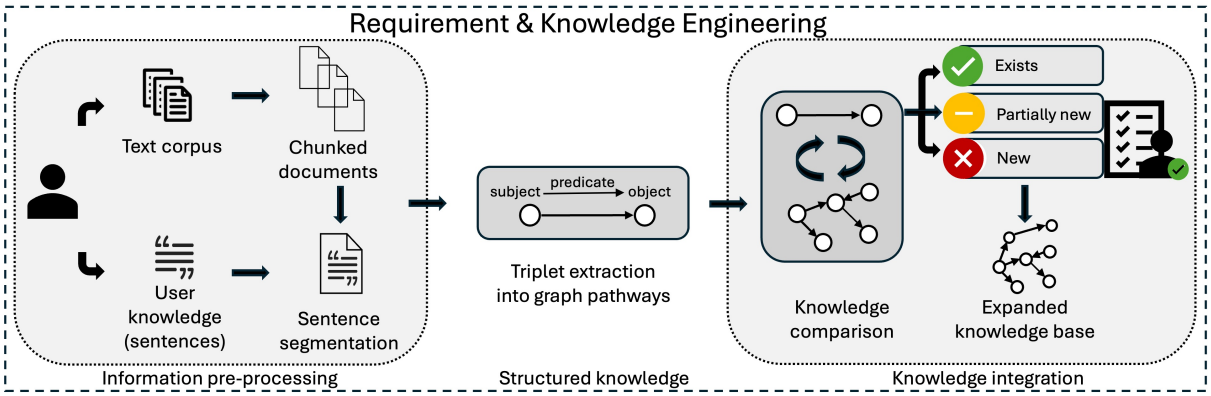

Figure 1: A flow diagram illustrating the process for extracting, structuring, and integrating tAI requirements into an expandable KB. Based on the extracted text and processed user input, the system generates new graph pathways, compares them to the existing knowledge, and uses human oversight to continuously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

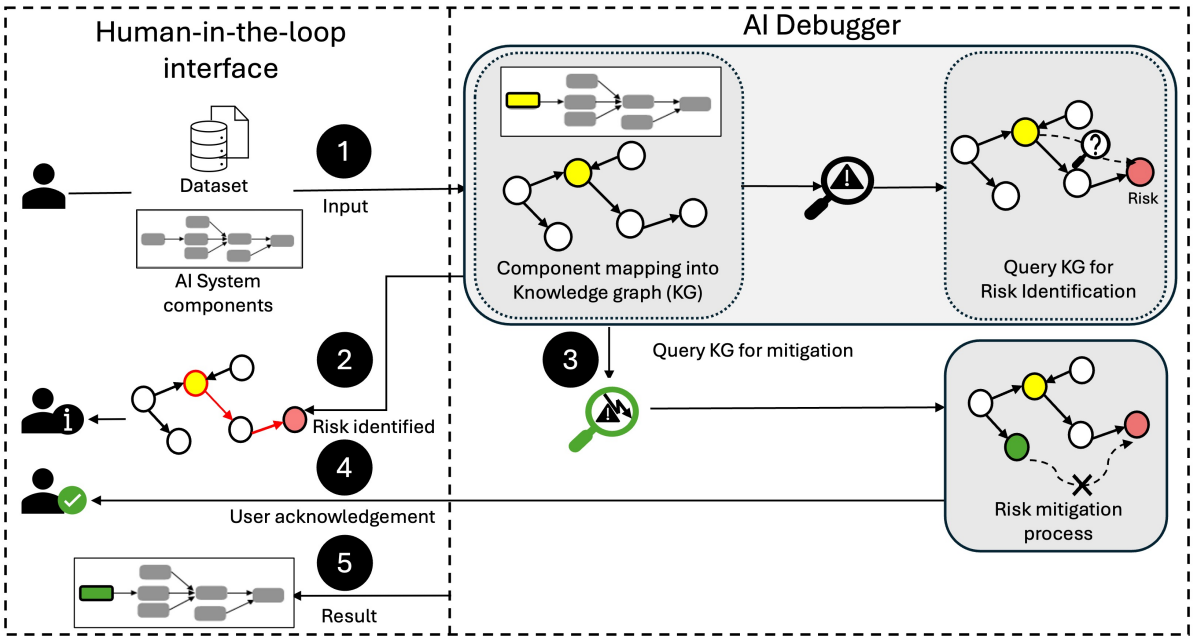

Figure 2: A flowchart depicting the **AI debugger** workflow. This process maps user inputs to the KG and query for potential risk identification and mitigation under HITL supervision.

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.