Trustworthy Requirements Generation for EU AI Act Compliance: A Knowledge Graph Approach

José Siqueira de Cerqueira[1]

[1]Tampere University, Tampere, Finland

Abstract

Although the EU AI Act establishes binding requirements for high-risk AI-based systems, developers lack the necessary tools to systematically derive legally grounded specifications from its 113 articles, 180 recitals and 13 annexes. While Large Language Models (LLMs) have the potential to bridge the gap between legal documents and software requirements, a lack of proper grounding could hinder their adoption. This PhD research project proposes TERE4AI (Trustworthy Ethical Requirements Engineering for AI), an open-source tool that generates requirements compliant with the EU AI Act from natural language system descriptions. TERE4AI combines a knowledge graph containing 590 semantic alignments between EU AI Act provisions and AI High-Level Expert Group (HLEG) ethical principles with a multi-agent pipeline aligned to classical requirements engineering phases. The tool implements six strategies to enhance trustworthiness, including knowledge graph grounding, multi-agent collaboration, and LLM-as-judge validation. These strategies allow developers to trust the generated outputs to a calibrated level, while maintaining oversight. We hope that the proposed tool will help developers build more ethically aligned and compliant AI-based systems, thus making AI ethics more practical.

Keywords

Requirements Engineering, EU AI Act, LLM Trustworthiness, Regulatory Compliance

1. Introduction

For years, the AI ethics community operated in what might be called the "soft law" era. Organizations published guidelines and principles—the AI HLEG Ethics Guidelines [1], the ECCOLA method [2], and dozens of similar frameworks [3]—yet these remained largely advisory. Developers faced challenges to translate abstract principles like "transparency" and "fairness" into concrete software specifications, and without legal consequences, compliance was often deprioritized.

The European Union's Artificial Intelligence Act [4] is changing this landscape by establishing the world's first legally binding framework for AI systems. However, the transition from abstract guidelines to binding regulation has not solved the core challenge facing developers. The EU AI Act spans 113 articles, 180 recitals, and 13 annexes, with complex interconnections and cross-references between provisions [4]. Developers typically lack legal expertise to interpret such extensive regulatory text [5], and SMEs face disproportionate compliance burdens—a survey of over 1,000 technology SMEs found that EU startups lose on average €94,000 to €322,000 due to regulatory compliance delays [6]. This creates a new bottleneck: regulatory requirements exist, but developers lack practical tool support to systematically derive actionable specifications from them.

The emergence of AI coding agents has intensified this need. Tools like Claude Code and GitHub Copilot have demonstrated remarkable capabilities in generating implementation code, but they require well-defined specifications to produce compliant systems. The challenge has shifted from "how do we write code?" to "how do we specify what the code should do to comply with regulations?"

Large Language Models (LLMs) offer a potential solution for bridging legal documents and software specifications. However, studies show that LLMs generate factually incorrect legal information between 58% and 88% of the time when answering legal queries without proper grounding [7]. This unreliability

Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2026. Poznań, Poland, March 23-26, 2026

jose.siqueiradecerqueira@tuni.fi (J. Siqueira de Cerqueira)

0000-0002-8143-1042 (J. Siqueira de Cerqueira)

demands trustworthiness-enhancing strategies—techniques such as knowledge graph grounding, multi-agent collaboration, and human-in-the-loop validation—that enable calibrated trust: allowing users to appropriately rely on AI outputs while maintaining oversight [8].

Problem Statement. Developers lack tool support to systematically derive legally-grounded requirements from extensive AI regulations, while LLMs without trustworthiness-enhancing strategies cannot reliably bridge legal documents and software specifications. This creates a bottleneck for building EU AI Act compliant systems, particularly affecting SMEs and startups without access to legal expertise.

This PhD project addresses this gap through two main contributions: (1) **TERE4AI**, an open-source tool that combines knowledge graph grounding with a multi-agent pipeline to generate EU AI Act compliant requirements with explicit legal citations and traceability; and (2) a **benchmark dataset** of compliance requirements across EU AI Act risk levels for evaluating requirement generation tools.

2. Related Work

Extensive research has been carried out on the challenges of translating ethical principles into practice. Jobin et al. [3] analyzed 84 AI ethics guidelines worldwide, finding convergence around five principles—transparency, fairness, non-maleficence, responsibility, and privacy—but significant divergence in implementation. Practical frameworks such as the ECCOLA method [2] provide card-based guidance for developers, yet these approaches require substantial human interpretation and lack automated tool support.

Requirements engineering has increasingly addressed regulatory compliance. Kosenkov et al. [9] conducted a systematic mapping study of 280 primary studies, arguing that compliance should be addressed early in software engineering—specifically during requirements engineering—rather than as an afterthought. However, existing RE approaches remain dispersed across domains and do not address the unique challenges of AI-specific regulations, particularly the EU AI Act's risk-based classification, interconnected cross-references, and mapping to ethical principles.

Regarding LLMs for RE tasks, Hemmat et al. [10] identified challenges including hallucination, domain knowledge gaps, and incomplete outputs. For legal compliance, Hassani et al. [11] argue that existing approaches rely on sentence-level analysis that ignores broader document context, and lack the ability to provide justification for compliance decisions. Dahl et al. [7] found that LLMs produce incorrect legal information 58–88% of the time, highlighting the need for grounding mechanisms.

To address LLM reliability, researchers have proposed trustworthiness-enhancing strategies. Our bibliometric analysis [8] identified 20 such strategies, including knowledge graphs, RAG, and multi-agent collaboration. In prior work [12], we found that combining multiple strategies improves output reliability. The organizational trust model by Mayer et al. [13]—conceptualizing trust through ability, benevolence, and integrity—has been applied to frame calibrated trust in AI outputs.

To the best of our knowledge, no existing solution spans from natural language system descriptions to legally-grounded requirements with explicit traceability to both EU AI Act articles and ethical principles. This PhD project addresses this gap by combining knowledge graph grounding with a multi-agent RE pipeline implementing multiple trustworthiness-enhancing strategies.

3. Methodology

This PhD project follows the Design Science Research (DSR) methodology [14], which is suitable for devising novel artifacts and evaluating them through iterations. The research addresses three research questions:

- **RQ1:** How can legislation and standards be structured into a knowledge graph to enable trustworthy LLM-based requirements generation?
- **RQ2:** To what extent does knowledge graph grounding improve the accuracy and coverage of LLM-generated compliance requirements compared to ungrounded approaches?

- **RQ3:** How does TERE4AI support appropriate trust calibration when developers use it for building EU AI Act compliant systems?

The exploration phase, which motivates the creation of TERE4AI and identifies design choices, is documented in previous studies [12, 8, 5]. Through this prior work, we identified a lack of systematic tool support for developers creating ethically aligned AI-based systems, and catalogued trustworthiness-enhancing strategies that could address LLM reliability limitations. Based on these gaps, we defined three design objectives for TERE4AI: (O1) automate risk classification according to EU AI Act categories; (O2) generate requirements with explicit legal citations and traceability; and (O3) implement multiple trustworthiness-enhancing strategies to support calibrated trust.

The tool has been demonstrated with example AI system descriptions covering all EU AI Act risk categories. For high-risk systems, TERE4AI generates requirements with explicit citations to EU AI Act articles and mappings to AI HLEG principles.

We plan a three-part evaluation strategy. First, a **baseline comparison study** where an LLM generates compliance requirements under two conditions: using only parametric knowledge versus with access to TERE4AI's knowledge graph. We will measure accuracy, coverage of applicable articles, and hallucination rates. Second, **expert validation** where a sample of generated requirements will be reviewed by legal and compliance experts to assess correctness of legal interpretations and validity of traceability links. Third, a **practitioner study** through semi-structured interviews and surveys with developers after hands-on use, assessing perceived usefulness, trust in outputs, and whether users appropriately calibrate their reliance based on confidence indicators. Throughout these evaluations, we will develop a benchmark dataset of compliance requirements across EU AI Act risk levels as a community resource.

4. Proposed Solution

TERE4AI (Trustworthy Ethical Requirements Engineering for AI) is an open-source web-based tool that generates legally-grounded requirements from natural language AI system descriptions.[1] Given a description of an AI system, TERE4AI produces a structured requirements report containing risk classification, applicable requirements with explicit legal citations, and traceability to ethical principles. Figure 1 illustrates the pipeline architecture: the tool takes natural language descriptions as input and produces requirements reports with explicit legal citations by leveraging a knowledge graph that semantically aligns EU AI Act provisions with AI HLEG ethical principles.

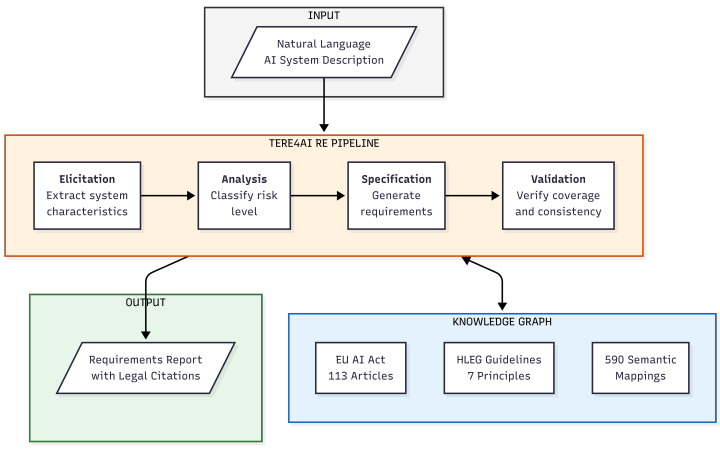

Figure 1: TERE4AI pipeline architecture.

The foundation of TERE4AI is a Neo4j knowledge graph containing the complete EU AI Act and AI HLEG Ethics Guidelines, along with their semantic alignments. The graph includes 113 articles, 519

[1]Available at https://github.com/josesiqueira/tere4ai

paragraphs, 375 points, 180 recitals, and 13 annexes from the EU AI Act, as well as 7 principles and 23 subtopics from the AI HLEG guidelines. We established 590 semantic alignments between EU AI Act paragraphs and AI HLEG principles using an LLM-based mapping process, where each paragraph was analyzed with contextual information and assigned relevance scores to applicable principles. The size and complexity of the EU AI Act—exceeding 100,000 tokens—prevents processing the entire regulation within LLM context windows. Our knowledge graph addresses this by enabling targeted retrieval of only the provisions relevant to each task, while preserving the interconnections and cross-references between articles that are essential for accurate compliance assessment.

We developed TERE4AI through four iterations. First, we constructed the knowledge graph from the EU AI Act, AI HLEG guidelines, and relevant ISO standards. Second, we implemented a Model Context Protocol (MCP) server to bridge the knowledge graph and LLM agents, addressing token limit constraints. Third, we developed the multi-agent pipeline with four specialized agents. Fourth, we implemented the web interface with confidence indicators. Each iteration provided feedback that guided subsequent development.

TERE4AI employs four sequential LLM-based agents, each aligned with a classical RE phase. The Elicitation Agent extracts structured system characteristics from natural language input, including domain, intended use, data types, and risk indicators such as fundamental rights impact, biometric processing, and vulnerable group involvement. The Analysis Agent classifies systems into EU AI Act risk categories through hierarchical decision logic: Article 5 prohibited practices yield unacceptable risk, Annex III categories yield high-risk, Article 50 transparency requirements yield limited risk, and remaining systems are classified as minimal risk. The Specification Agent generates requirements for non-prohibited systems, with each requirement including a formal statement, EU AI Act citation with quoted text, mapped AI HLEG principle with relevance score, and verification criteria. The Validation Agent verifies completeness by checking article coverage and AI HLEG principle coverage, and detects conflicts among generated requirements. The MCP server provides a semantic abstraction layer between agents and the knowledge graph, and also exposes the knowledge graph to external MCP-compatible coding agents, extending TERE4AI's legal grounding to broader development workflows.

TERE4AI implements six trustworthiness-enhancing strategies identified in our bibliometric analysis [8]: multi-agent collaboration distributes reasoning across specialized agents; structured communication enforces schema validation through strongly-typed models at each pipeline stage; knowledge graph grounding ensures requirements derive from retrieved legal text; self-assessment via LLM-as-judge allows the Validation Agent to critically assess outputs; coverage metrics quantify article and principle coverage as percentages; and source traceability ensures every requirement includes explicit citations. These strategies collectively support calibrated trust—enabling users to appropriately rely on outputs while maintaining critical oversight.

This PhD project makes two main contributions. First, TERE4AI represents a novel integration of knowledge graph grounding with a multi-agent RE pipeline for regulatory compliance—spanning from natural language system descriptions to formally-specified requirements with explicit traceability to both legal articles and ethical principles. Second, the benchmark dataset of compliance requirements across EU AI Act risk levels will provide a reproducible evaluation resource for the research community. The novelty lies in the semantic alignment between binding legal provisions and ethical guidelines, the application of multiple trustworthiness-enhancing strategies in combination, and the focus on calibrated trust through confidence indicators and coverage metrics.

5. Research Progress and Future Work

This PhD project started in September 2023 and is expected to conclude in August 2027, with approximately 18 months remaining. The research has progressed through several phases, each building upon the previous to arrive at the current state of TERE4AI.

The first year focused on understanding the landscape of AI ethics operationalization and the challenges developers face in building compliant systems. This exploration resulted in a position paper

analyzing the EU AI Act's limitations from a developer perspective [5]. The second year investigated trustworthiness challenges in LLM-based systems, leading to two studies: an empirical evaluation of multi-agent collaboration strategies for ethical AI assessment [12], and a bibliometric analysis that identified 20 trustworthiness-enhancing strategies across the literature [8]. These studies established the theoretical foundation for TERE4AI's design.

The third year has focused on the design and implementation of TERE4AI itself. We constructed the knowledge graph containing the EU AI Act, AI HLEG guidelines, and 590 semantic alignments between them. We developed the multi-agent pipeline with four specialized agents aligned to RE phases, implemented the MCP server for knowledge graph access, and built the web interface with confidence indicators. A tool demonstration paper has been submitted describing the initial implementation.

The remaining 18 months will focus on evaluation and dissemination. The immediate next steps include conducting the baseline comparison study to measure the impact of knowledge graph grounding on requirement generation accuracy. This will be followed by recruiting legal experts for validation of generated requirements, and practitioners for the trust calibration study. In parallel, we will develop the benchmark dataset of compliance requirements across EU AI Act risk levels.

Several challenges lie ahead. Recruiting legal experts with EU AI Act knowledge for validation may prove difficult given the regulation's novelty. The practitioner study requires participants with both AI development experience and compliance awareness, a relatively specialized profile. Additionally, the EU AI Act implementation timeline means that real-world compliance practices are still emerging, which may affect the ecological validity of our evaluation. Finally, maintaining the knowledge graph as standards evolve and new guidance is published will require ongoing effort beyond the PhD timeline.

The expected outputs from the remaining work include: a journal paper on the knowledge graph construction and semantic alignment methodology (RQ1), a conference paper on the empirical evaluation comparing grounded versus ungrounded requirement generation (RQ2), a journal paper on practitioner trust calibration with LLM-based compliance tools (RQ3), and the open benchmark dataset as a community resource.

6. Conclusion

This paper presents ongoing PhD research addressing the challenge of deriving legally-grounded requirements from extensive AI regulations. We propose TERE4AI, an open-source tool that combines knowledge graph grounding with a multi-agent pipeline to generate EU AI Act compliant requirements with explicit legal citations and traceability to ethical principles. By implementing multiple trustworthiness-enhancing strategies, TERE4AI aims to support calibrated trust—enabling developers to appropriately rely on generated outputs while maintaining critical oversight.

The research contributes to both the requirements engineering and AI ethics communities: TERE4AI provides practical tool support for compliance, while the planned benchmark dataset will enable reproducible evaluation of requirement generation approaches. With 18 months remaining, the focus shifts to empirical evaluation through baseline comparisons, expert validation, and practitioner studies. We hope this work helps bridge the gap between binding AI regulations and the developers who must implement compliant systems.

Acknowledgments

This research was supported by CONVERGENCE of Humans and Machines (220025) and the EVIL-AI "The identification and the mitigation of the negative effects of Artificial Intelligence Agents" (JAES/2024/EVIL-AI) projects by Jane and Aatos Erkko Foundation and the "Multifaceted ripple effects and limitations of human-AI interplay at work, business and society (SYNTHETICA)" project (358714) by Research Council of Finland.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: Paraphrase and reword; Improve writing style; Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[2] V. Vakkuri, K.-K. Kemell, P. Abrahamsson, Eccola—a method for implementing ethically aligned ai systems, in: Journal of Systems and Software, volume 182, Elsevier, 2021, p. 111067. doi:10.1016/j.jss.2021.111067.

[3] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. doi:10.1038/s42256-019-0088-2.

[4] European Parliament and Council, Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[5] C. V. Sillberg, J. S. De Cerqueira, P. Sillberg, K.-K. Kemell, P. Abrahamsson, The eu ai act is a good start but falls short, in: International Conference on Software Business, Springer, 2024, pp. 114–130.

[6] ACT | The App Association, The Hidden Cost of AI Regulations: A Survey of EU, UK, and U.S. Companies, Technical Report, ACT | The App Association, 2025. URL: https://actonline.org/the-hidden-cost-of-ai-regulations-a-survey-of-eu-uk-and-u-s-companies/.

[7] M. Dahl, V. Magesh, M. Suzgun, D. E. Ho, Large legal fictions: Profiling legal hallucinations in large language models, Journal of Legal Analysis 16 (2024) 64–93. doi:10.1093/jla/laae003.

[8] J. S. de Cerqueira, K.-K. Kemell, R. Rousi, N. Xi, J. Hamari, P. Abrahamsson, Mapping trustworthiness in large language models: A bibliometric analysis bridging theory to practice, arXiv preprint arXiv:2503.04785 (2025).

[9] O. Kosenkov, P. Elahidoost, T. Gorschek, J. Fischbach, D. Mendez, M. Unterkalmsteiner, D. Fucci, R. Mohanani, Systematic mapping study on requirements engineering for regulatory compliance of software systems, Information and Software Technology 178 (2025) 107622. doi:10.1016/j.infsof.2024.107622.

[10] H. K. Hemmat, P. R. Anish, A. Ferrari, M. Abbas, Research directions for using large language models in software requirements engineering, Frontiers in Computer Science 7 (2025) 1514488. doi:10.3389/fcomp.2025.1514488.

[11] S. Hassani, M. Sabetzadeh, D. Amyot, J. Liao, Rethinking legal compliance automation: Opportunities with large language models, in: 2024 IEEE 32nd International Requirements Engineering Conference (RE), IEEE, 2024, pp. 432–440. doi:10.1109/RE59067.2024.00055.

[12] J. A. S. de Cerqueira, M. Agbese, R. Rousi, N. Xi, J. Hamari, P. Abrahamsson, Can we trust ai agents? a case study of an llm-based multi-agent system for ethical ai, arXiv preprint arXiv:2411.08881 (2024).

[13] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of Management Review 20 (1995) 709–734. doi:10.5465/amr.1995.9508080335.

[14] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Quarterly 28 (2004) 75–105.