Trustworthy LLMs for Ethically Aligned AI-based Systems: A PhD Research Plan

José Antonio Siqueira de Cerqueira[1,*], Rebekah Rousi[2], Nannan Xi[1], Juho Hamari[1], Kai-Kristian Kemell[1] and Pekka Abrahamsson[1]

[1]*Tampere University (TAU), Finland*
[2]*University of Vaasa (UWASA), Finland*

Abstract

In response to growing concerns around trustworthiness and ethical alignment in AI systems, this PhD aims to investigate how Large Language Models (LLMs) can be leveraged to support ethically aligned AI development in software engineering. Despite advancements, integrating ethical principles into AI workflows remains challenging, particularly in real-world applications that require compliance with emerging regulations, such as the EU AI Act. We will develop a Visual Studio Code (VSCode) Generative AI (GenAI) Extension powered by a multi-agent LLM system with Retrieval-Augmented Generation (RAG) capabilities. The extension will be designed to aid developers by evaluating code compliance with ethical standards, providing actionable recommendations to embed trustworthiness from early stages of development. The GenAI Extension will be evaluated through an iterative design science approach, encompassing dataset generation, ethical benchmarking, and practitioner testing. A dataset of over 2000 ethically aligned AI systems, will be created in compliance with leading regulatory frameworks, serving as a foundation for this tool's assessments. With this work, we hope to assist developers, particularly in startups and SMEs, by providing practical resources for building ethically aligned AI within limited resources. Through this approach, we aim to bridge the gap between abstract ethical principles and actionable software development practices, making ethical AI more accessible across industry contexts.

Keywords

AI ethics, Large Language Models, Trustworthiness, AI4SE

1. Problem Definition

In today's increasingly digitized world, Artificial Intelligence (AI) is emerging as a transformative force, reshaping industries, economies, and daily lives. From virtual assistants and recommendation algorithms to autonomous vehicles and medical diagnostics. AI-based systems, particularly Large Language Models (LLMs), are becoming ubiquitous, wielding considerable influence over decision-making processes and human interactions [1, 2]. LLM is a subfield of AI developed through the use of complex algorithms and large amounts of data [1, 2]. It is permeating every area of science and in people's everyday lives [3]. However, many reports reveal that its use – or misuse – can cause significant harm, directly or indirectly [2]. For example, it can produce factual inaccuracies, provide biased information, hallucinations, racism and mysoginism [1, 4]. This is largely due to the nature of LLMs, which reproduces patterns found in the data on which it has been trained on [2]. Furthermore, the algorithms that generate each word are probabilistic. In other words, the last word generated depends on the probability of its occurrence depending on the preceding word [5]. As a result, they are untrustworthy by nature, that is, despite generating coherent text, LLMs operate without genuine understanding, leading to outputs that may be irrelevant or misleading [5]. These discussions are crucial as our reliance on LLMs for tasks and decision-making grows, especially in software engineering [3]. In the Software Engineering field, the capabilities of LLMs are being explored in the software development, maintenance, and evolution [6, 7, 8]. Accordingly, they find applications across various stages of the software development process,

The 15th International Conference on Software Business (ICSOB 2024), November 18–20, 2024, Utrecht, The Netherlands
*Corresponding author.
jose.siqueiradecerqueira@tuni.fi (J. A. S. d. Cerqueira); rebekah.rousi@uwasa.fi (R. Rousi); nannan.xi@tuni.fi (N. Xi); juho.hamari@tuni.fi (J. Hamari); kai-kristian.kemell@tuni.fi (K. Kemell); pekka.abrahamsson@tuni.fi (P. Abrahamsson)
0000-0002-8143-1042 (J. A. S. d. Cerqueira); 0000-0001-5771-3528 (R. Rousi); 0000-0002-9424-8116 (N. Xi); 0000-0002-6573-588X (J. Hamari); 0000-0002-0225-4560 (K. Kemell); 0000-0002-4360-2226 (P. Abrahamsson)

including requirement analysis, software design, code implementation, testing, refactoring, defect detection, and repair [7].

Regarding AI, it has been noted for several years that it faces ethical problems, similar to those faced by LLMs [9, 10]. However, researchers and the industry have approached AI ethics in a more theoretical way, providing abstract ethical guidelines and principles [11]. Recent advances in legislation, such as the EU AI Act, propose to regulate the development and use of AI-based systems [12], but there is still no evidence of the extent to which it can assist practitioners in operationalising AI ethics. Therefore, there is a problem in bridging the gap between theory and practice in AI ethics, as well as in addressing the trustworthiness of LLMs.

2. Knowledge Gap

Regarding trustworthiness in LLMs, efforts found in the literature focus on finding a taxonomy with trustworthiness aspects, e.g., truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, accountability, regulations and law [4]. Moreover, taxonomies serve as a way on how to assess LLMs in relation to trustworthiness. Having specialized roles [13, 14, 15], the use of external tools (e.g., running a code, searching the web) [13, 15], providing human interaction (i.e., human in the loop) [15], structured conversations (i.e., message templates) [13, 15] and different conversational patterns [15], are pointed out as techniques to improve overall trustworthiness of LLM systems. These techniques can significantly improve their reasoning as they debate and refine their discourse over multiple rounds. However, introducing new layers of complexity and challenges, such as increasing the overall cost (requires multiple instances and rounds) [16] and scalability (manage computational resources) [17]; while this approach is innovative, it is still generative AI, so it can produce convincing but wrong results [16], generates a different software on each run [14] and is prone to possible unintentional harmful outcomes and vulnerable to misuse [14]. Similarly to trustworthiness in LLMs, AI ethics also lack a centralized set of principles, assessment and practical guidance.

While recent interest from academia and industry highlights AI ethics as a growing field of research, concerns regarding AI's development and deployment have a long history, with several incidents drawing public attention recently [18]. As a response, multiple principles and guidelines have been formulated in recent years by diverse stakeholders, including academia, industry, and civil society, to delineate what constitutes ethical AI [10]. Ryan and Stahl [19] identified 11 foundational ethical principles relevant to AI ethics: 1) Transparency, 2) Justice and Fairness, 3) Non-maleficence, 4) Responsibility, 5) Privacy, 6) Beneficence, 7) Freedom and Autonomy, 8) Trust, 9) Sustainability, 10) Dignity, and 11) Solidarity. Nonetheless, AI ethics is still an open debate, where practitioners are often disoriented with abstract principles, lacking clear guidance on how to operationalise the many ethical principles available [18].

The European Parliament is progressing with the world's first AI regulation [20], underlining the current status of most guidelines as "soft law," without mandatory enforcement or significant legal repercussions [21]. This regulation echoes the abstract nature with which AI ethics is typically approached [18, 10, 22]. Challenges in translating these broad ethical principles into actionable practices stem from the subjective interpretation required by practitioners to apply them in real-world scenarios [21]. Despite its critical role, ethical considerations in AI design and implementation are often addressed only later in the development process [23].

In the literature, several studies emphasize that examining trustworthiness in LLMs requires situational applications, where models are tested within specific contexts to effectively assess how trustworthiness issues unfold and address unique challenges [4, 24]. We argue that an appealing emergent application of LLM agents is in the development of ethically aligned AI-based systems. Concerning the use of LLM in Software Engineering (LLM4SE), practitioners should trust the solutions, and they must be seamlessly adopted by practitioners, otherwise they can become barriers [25]. For the best of our knowledge, there are no studies that directly address the development of ethically aligned AI-based systems through the use of LLMs. Unlike prior approaches in the literature, this work explores the

application of LLM-based multi-agent systems in AI development, emphasizing the incorporation of ethical principles from the earliest stages of the development lifecycle.

3. Research Method

To address the gaps identified in the literature, we aim to create a Visual Studio Code (VSCode) Generative AI (GenAI) Extension tool. VSCode is widely used in the industry, with approximately 75% of developers reporting it as their preferred code editor in the 2023 Stack Overflow Developer Survey [26]. This GenAI Extension will assist developers build ethically aligned AI-based systems by assessing the code and suggesting possible code. Nevertheless, we will follow some steps for the creation of this tool, following the Design Science Research method to build and evaluate IS artefacts [27].

Firstly, we will identify techniques to improve trustworthiness in LLMs and develop a prototype, the LLM-based multi-agent system with Retrieval Augmented Generation (RAG). Next, we will benchmark our prototype against the SWE-benchmark. This will be done to test the accuracy and trustworthiness of our system. After that, we will create a dataset of more than 2000 ethically aligned AI-based systems generated by using the system and complying with new legislation addressing AI ethics. This will be done by using the AI Incidents Database and by feeding the (1) EU AI Act, (2) AI HLEG, (3) ISO-IEC 42001:2024 and (4) California's GenAI bills, into our system. Then, the dataset will be used to create a novel benchmark, to assess other LLMs regarding their capability to generate ethically aligned AI-based system. Finally, we will create our VSCode GenAI Extension tool and test it with practitioners in terms of synergy and trust [25].

3.1. Research Questions

The following research questions guide this study:

- **RQ1:** What techniques can be identified and applied to enhance the trustworthiness of LLM-based systems in software engineering (LLM4SE)?
- **RQ2:** How can LLMs be utilized to evaluate and develop AI systems that are ethically compliant with the EU AI Act?
- **RQ3:** How does VSCode GenAI extension influence synergy, trust, and ethical AI development outcomes in startups and SMES?

4. Timeline

Phase	Description and Milestones
Sep 2023 – Febr 2024: Foundational Research and Exploration	- Conduct an in-depth review of the EU AI Act to identify regulatory standards for ethically aligned AI systems. - Explore existing techniques to enhance trustworthiness in LLMs, focusing on ethical guidelines and principles. - Collect resources such as the AI Incident Database and international ethics standards.
Mar 2024 – Nov 2024: Prototype Design and Early Development	- Define key trustworthiness and ethical alignment criteria specific to LLMs.

	- Design and prototype an LLM-based multi-agent system leveraging Retrieval-Augmented Generation (RAG) capabilities.
Dec 2024 – Mar 2025: Prototype Refinement and Benchmark Setup for LLM4SE	- Perform necessary refinements to the prototype and conduct benchmarks in LLM4SE using the SWE-benchmark.
Apr 2025 – Feb 2026: Dataset Generation and Novel Benchmark Creation	- Generate a dataset of over 2000 ethically aligned AI-based systems using the prototype. - Develop a novel benchmark to assess other LLMs' abilities to generate ethically aligned systems, based on insights from the AI Incident Database and legislative sources.
Mar 2026 – Aug 2026: VSCode GenAI Extension Development	- Create the GenAI Extension tool for Visual Studio Code (VSCode), incorporating trustworthiness and ethical compliance assessment functionalities. - Conduct initial usability testing with developers to refine features based on synergy and trust insights.
Sep 2026 – Feb 2027: Extension Refinement and Practitioner Testing	- Refine the extension based on developer feedback, focusing on enhanced trustworthiness and ethical adherence. - Conduct extensive testing with practitioners to validate usability and ethical compliance in real-world applications.
Mar 2027 – Aug 2027: Final Adjustments and Knowledge Dissemination	- Implement final adjustments for industry readiness of the VSCode extension. - Publish and present findings on trustworthiness in LLMs, ethical compliance, and practical implications in software engineering.

Table 1: Project Timeline

5. Preliminary Results

Here we will present some of our preliminary results, with an initial prototype using OpenAI API gpt-4o that relies only on internal knowledge, that is, without RAG. This prototype, called LLM-based multi-agent system (LLM-BMAS) was developed by implementing different techniques to improve

trustworthiness in AI for software engineering (AI4SE), and evaluated against three real AI incidents found in AI Incidents Database [28]. The evaluation was done using thematic analysis, hierarchical clustering, ablation study, and source code execution. Our initial results show that LLM-BMAS has the ability to provide extensive and detailed source code and documentation, around 2,000 lines, while ablation study - using only ChatGPT user interface as baseline - produce around 80 lines without source code. Moreover, it is seen from the thematic analysis and hierarchical clustering that the prototype can address various ethical issues in AI that are often overlooked, e.g., bias, transparency, fairness.

However, several factors currently impede seamless integration for practitioners [25]. Notably, these challenges include limited practicality in extracting source code from generated text—especially when handling complex modules—as well as difficulties with installing packages and managing outdated dependencies tied to the model's original training date. Although advancements can enhance the trustworthiness and quality of LLM4SE applications, further improvements are essential to enhance practical usability for developers.

6. Expected Contributions

This research aims to contribute to the field of software engineering by developing a novel Visual Studio Code (VSCode) GenAI Extension that integrates LLM-based multi-agent systems to support the creation of ethically aligned AI systems. The extension will incorporate trustworthiness assessments based on a unique dataset of over 2000 AI-based systems that align with key regulatory frameworks such as the EU AI Act, AI HLEG guideline, and ISO-IEC 42001:2024. By establishing new benchmarks specific to ethical AI development, this tool will enable developers to assess and enhance code compliance with ethical standards early in the development process. The result is expected to bridge the gap between theoretical ethics principles and practical application in software engineering.

In addition, this project aims to advance practical trustworthiness techniques for LLMs in Software Engineering (LLM4SE). Rigorous testing with software practitioners will evaluate the effectiveness of the extension in providing ethically guided code recommendations, focusing on usability, trust and real-world synergy. By providing a structured and accessible approach to embedding ethical principles into standard development practices, this work can particularly support practitioners in start-ups and small to medium enterprises where resources and regulatory expertise may be limited. This contribution is expected to make ethical AI development more feasible for smaller teams, helping them to align their AI systems with evolving regulatory and ethical standards from the earliest stages of development.

Acknowledgments

This research was supported by Jane and Aatos Erkko Foundation through CONVERGENCE of Humans and Machines Project under grant No. 220025.

Declaration on Generative AI

During the preparation of this work, the authors utilized ChatGPT to assist in identifying and correcting writing errors, and enhancing clarity and conciseness. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment, arXiv preprint arXiv:2308.05374 (2023).

[2] P. P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. doi:10.1016/j.iotcps.2023.04.003.
[3] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023).
[4] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, et al., TrustLLM: Trustworthiness in large language models, arXiv preprint arXiv:2401.05561 (2024).
[5] L. Floridi, M. Chiriatti, GPT-3: its nature, scope, limits, and consequences, Minds Mach. 30 (2020) 681–694. doi:10.1007/S11023-020-09548-1.
[6] I. Ozkaya, Application of large language models to software engineering tasks: Opportunities, risks, and implications, IEEE Software 40 (2023) 4–8. doi:10.1109/MS.2023.3248401.
[7] X. Peng, Software development in the age of intelligence: embracing large language models with the right approach, Frontiers of Information Technology & Electronic Engineering 24 (2023) 1513–1519. doi:10.1631/FITEE.2300537.
[8] B. Ni, M. J. Buehler, Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge, Extreme Mechanics Letters 67 (2024) 102131. doi:10.1016/j.eml.2024.102131.
[9] T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, Minds and Machines 30 (2020) 99–120. doi:10.1007/s11023-020-09517-8.
[10] 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.
[11] V. Vakkuri, K.-K. Kemell, P. Abrahamsson, ECCOLA–a method for implementing ethically aligned ai systems, arXiv preprint arXiv:2004.08377 (2020).
[12] A. R. Marinković, The new EU AI Act: A comprehensive legislation on AI or just a beginning?, Global Journal of Business and Integral Security (2023).
[13] S. Hong, X. Zheng, J. P. Chen, Y. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, Z. H. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, MetaGPT: Meta programming for multi-agent collaborative framework, ArXiv abs/2308.00352 (2023).
[14] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun, Communicative agents for software development, arXiv preprint arXiv:2307.07924 (2023).
[15] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, AutoGen: Enabling next-gen LLM applications via multi-agent conversation framework, arXiv preprint arXiv:2308.08155 (2023).
[16] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).
[17] Y. Talebirad, A. Nadiri, Multi-agent collaboration: Harnessing the power of intelligent LLM agents, arXiv preprint arXiv:2306.03314 (2023).
[18] E. Halme, M. Jantunen, V. Vakkuri, K. Kemell, P. Abrahamsson, Making ethics practical: User stories as a way of implementing ethical consideration in software engineering, Inf. Softw. Technol. 167 (2024) 107379. doi:10.1016/J.INFSOF.2023.107379.
[19] M. Ryan, B. C. Stahl, Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications, Journal of Information, Communication and Ethics in Society 19 (2021) 61–86. doi:10.1108/JICES-12-2019-0138.
[20] E. Commission, EU AI Act: First regulation on artificial intelligence, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. Accessed 01 Apr 2024.
[21] J. A. S. de Cerqueira, A. P. D. Azevedo, H. A. T. Leão, E. D. Canedo, Guide for artificial intelligence ethical requirements elicitation - RE4AI ethical guide, in: 55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022, ScholarSpace, 2022, pp. 1–10. URL: http://hdl.handle.net/10125/80015.
[22] N. K. Corrêa, C. Galvão, J. W. Santos, C. D. Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem, N. de Oliveira, Worldwide AI ethics: A review of 200 guidelines and recommen-

dations for AI governance, Patterns 4 (2023) 100857. doi:10.1016/J.PATTER.2023.100857.

[23] V. Vakkuri, K. Kemell, M. Jantunen, E. Halme, P. Abrahamsson, ECCOLA - A method for implementing ethically aligned AI systems, J. Syst. Softw. 182 (2021) 111067. doi:10.1016/J.JSS.2021.111067.

[24] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, DecodingTrust: A comprehensive assessment of trustworthiness in GPT models, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023. doi:10.48550/arXiv.2306.11698.

[25] D. Lo, Trustworthy and synergistic artificial intelligence for software engineering: Vision and roadmaps, in: IEEE/ACM International Conference on Software Engineering: Future of Software Engineering, ICSE-FoSE 2023, Melbourne, Australia, May 14-20, 2023, IEEE, 2023, pp. 69–85. doi:10.1109/ICSE-FOSE59343.2023.00010.

[26] Stack Overflow Developer Survey 2023, https://survey.stackoverflow.co/2023/, 2023. Accessed 25 Oct 2024.

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

[28] J. A. S. de Cerqueira, M. Agbese, R. Rousi, N. Xi, J. Hamari, P. Abrahamsson, Can we trust AI agents? An experimental study towards trustworthy LLM-based multi-agent systems for AI ethics, arXiv preprint arXiv:2411.08881 (2024).