Ayman Asad Khan

"CAN SOFTWARE DEVELOPMENT ADOPT AI ETHICS?"

Agents4EthicalSE: A Multi-Agent RAG Framework

Master's Thesis
Faculty of Information Technology and Communication Sciences (ITC)
Examiner: José Antonio Siqueira de Cerqueira
Examiner: Kai-Kristian Kemell
July 2025

ABSTRACT

Ayman Asad Khan: CAN SOFTWARE DEVELOPMENT ADOPT AI ETHICS?
Agents4EthicalSE: A Multi-Agent RAG Framework
Master's Thesis
Tampere University
Master's Degree Programme in Software, Web & Cloud
July 2025

As Large Language Models (LLMs) based solutions such as GitHub Copilot, Amazon CodeWhisperer, Rubberduck (ChatGPT based) become widely used in software development, questions remain about how these tools can support ethical and regulatory compliance. While existing AI ethics guidelines outline high-level principles such as fairness, transparency, and accountability, developers lack tools that operationalize these during day-to-day coding tasks. This thesis introduces *Agents4EthicalSE*, a multi-agent Retrieval-Augmented Generation (RAG) framework that assists developers in generating ethically grounded code aligned with the EU AI Act, AI HLEG guidelines, and other relevant regulatory documents.

The framework is designed as a practical alternative to baseline code generation workflows. In a typical setup, tools like Copilot produce outputs without referencing legal or ethical constraints. In contrast, *Agents4EthicalSE* consists of specialized agents—such as a RiskGuard and an AI Ethicist—that engage in structured multi-round interactions. The AI Ethicist uses RAG to retrieve relevant content from regulatory texts and guidelines, enabling real-time critique and refinement of code suggestions. Developers can review these outputs and iterate with agent feedback to address ethical risks or non-compliance concerns.

We evaluated the system through a workshop-based study involving 82 AI practitioners and developers. Participants were asked to simulate development tasks using both *Agents4EthicalSE* and a baseline LLM. The tasks were derived from real-world AI failures documented in the AI Incident Database, reframed as project descriptions across different risk categories defined by the EU AI Act. Survey results showed that *Agents4EthicalSE* helped participants better identify ethically sensitive components, interpret compliance requirements, and reflect on ethical trade-offs during code generation. Feedback also pointed to areas for improvement, such as citation clarity and interface responsiveness.

These results suggest that multi-agent RAG frameworks can provide developers with more interpretable and regulation-aware coding assistance compared to standard LLMs. This work contributes a step toward integrating ethical reasoning into software engineering workflows through structured, document-informed interactions.

Keywords: AI Ethics, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Large Language Models (LLM), Responsible AI, AI4SE.

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

☐ No
☒ Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

1. Quadratic – for data analysis and visualization
2. ChatGPT – for brainstorming, drafting, and refining text
3. OpenAI Assistants API – for building and coordinating specialized agents in the Agents4EthicalSE framework
4. OpenAI Vector Store – for implementing document retrieval in the RAG pipeline
5. **GPT-4o model** – as the core large language model (LLM) powering agent responses and code generation in the experimental tool

Purpose of using AI tools: AI tools were used to support both the **research process** and the **tool development**.

- *Quadratic* was used to analyze and visualize participant responses in the evaluation study.
- *ChatGPT* assisted in brainstorming ethical scenarios, refining technical descriptions, and structuring content (especially during literature synthesis and prompt design).
- The *OpenAI Assistants API* enabled the creation of multi-agent conversations with predefined roles (e.g., RiskGuard, AI Ethicist) and structured dialogue.
- *OpenAI Vector Store* was employed in indexing regulatory documents (e.g., EU AI Act) and enabling RAG-based retrieval.
- *GPT-4o* served as the LLM engine in the Agents4EthicalSE prototype, producing responses, generating Python code, and grounding recommendations in retrieved content.

Sections where AI tools were used:

- **Section 3.3.3 (Technology Stack and Implementation)** – for integrating GPT-4o, Assistants API, and Vector Store
- **Section 4 (Evaluation and Results)** – for analyzing workshop survey data using Quadratic.
- **Sections 1 and 2** – for drafting and refining conceptual discussions using ChatGPT.
- **Appendices** – for processing and organizing references and literature mapping.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

PREFACE

This thesis was written as part of my Master of Science studies in the Faculty of Information Technology and Communication Sciences at Tampere University. The research was conducted between August 2024 and April 2025 focusing on the integration of AI ethics into software development through multi-agent large language model systems.

I would like to express my sincere gratitude to my supervisor, **José Siqueira de Cerqueira**, for introducing me to the domain of AI ethics—not as a theoretical concern, but as a critical lens for responsible software engineering. His belief that ethics must not be treated as an afterthought, but instead embedded directly into the development process, deeply shaped the foundation of this work. I am especially thankful for his guidance, constructive feedback, and continuous support throughout the research process. I also want to thank **Dr. Kai-Kristian Kemell** for his comments, which helped refine the clarity, structure, and academic rigor of this thesis.

I would like to thank the practitioners who participated in the user study, providing feedback and reflections that were essential for evaluating the proposed framework. Special thanks to the research group and colleagues at GPT-Lab, Tampere University for their input and encouragement.

Finally, I am grateful to my family and friends for their patience and support during the thesis journey.

Tampere, Finland, 12 May 2025

Ayman Asad Khan

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence (European Commission)
AI SE	AI in Software Engineering
AI4SE	Artificial Intelligence for Software Engineering
API	Application Programming Interface
DSR	Design Science Research
EU	European Union
EU AI Act	European Union Artificial Intelligence Act
GPT-4o	Generative Pre-trained Transformer 4 Omni
HLEG	High-Level Expert Group
IDE	Integrated Development Environment
LLM	Large Language Model
MAS	Multi-Agent System
NLP	Natural Language Processing
OpenAI	Artificial intelligence research company providing LLMs (e.g., ChatGPT, GPT-4o)
RAG	Retrieval-Augmented Generation
RE4AI	Requirements Engineering for AI Ethics
RQ	Research Question
SE	Software Engineering
SLR	Systematic Literature Review

INTRODUCTION

1.1 Background and Motivation

Artificial intelligence (AI) systems are becoming common in software products and services across various domains, from healthcare to finance. This advancement has heightened concerns about how to **operationalize AI ethics** in software development practice. Over the past few years, numerous high-level AI ethics frameworks and principle-based guidelines have been published by industry consortia, governments, and academia [1][2]. In fact, by mid-2019 over 80 such ethics guides existed in the public domain [1]. These documents typically articulate principles that AI systems should uphold, for example, fairness, transparency, accountability, and privacy. Initially, this principles-based movement was seen as a defining step towards ensuring ethical AI development [3][4]. However, **translating these abstract principles into concrete software design and implementation practices has proven difficult**. Even as ethical AI guidelines multiplied, real-world incidents of AI misuse or bias continued to emerge in areas like healthcare, education, and law enforcement [4]. This paradox highlights the **gap between "principles" and "practice"** in AI ethics [4][5]. In other words, having ethical principles on paper is necessary but far from sufficient for ensuring ethical outcomes in software systems.

One major limitation of existing **principles-based frameworks** is their level of abstraction. AI ethics guidelines often remain high-level and lack actionable guidance for software developers and engineers [4]. As a result, developers may agree with ethical values in theory but struggle to apply them during day-to-day development tasks. This challenge can lead to *"ethics shopping" or "ethics washing"*, where organizations selectively adopt convenient guidelines or treat ethics as a PR checkbox without changing core practices [2][4]. The result is that many AI products continue to exhibit issues like unfair bias or lack of transparency despite the proliferation of ethical codes. This situation has caused growing concern among researchers and practitioners about how to genuinely embed ethical considerations into the software development lifecycle, rather than simply formulating principles at a policy level [6].

From a software engineering perspective, **operationalizing AI ethics** means incorporating ethical requirements and checks into the same processes that govern functional and business requirements. Recent studies indicate this integration is nascent at best. For instance, Agbese *et al.* [7] interviewed software engineering executives and found that they primarily treated certain ethical concerns, such as privacy, merely as legal compliance issues, with *"no other consideration for ethical requirements identified"* beyond what the law mandates. In that study, apart from privacy and data protection (seen as GDPR compliance), other ethical principles (like fairness, transparency, or societal well-being) were absent from software project management practices. This finding underscores the **limitations of current practice**: practitioners require practical **tools and methodologies** that do not demand deep expertise in philosophy and that fit naturally into existing development workflows [8]. In summary, the motivation for this thesis stems from a recognized need in both the AI ethics literature and software engineering community: to find effective ways of **embedding ethical reasoning and compliance into the software development lifecycle**. The question is how to achieve this in practice, given that purely principle-driven approaches have had limited impact.

Against this backdrop, emerging advances in AI technology itself may offer **new opportunities to operationalize ethics**. In particular, the evolution of powerful *large language models (LLMs)* and their ability to function as conversational agents has opened possibilities for ethical deliberation within software tools. Traditional software engineering relies on human stakeholders (developers, ethicists, regulators) to discuss and enforce ethical guidelines. However, AI agents can also assist or simulate parts of this process. **Multi-agent LLM systems** – where multiple AI agents with separate roles communicate with each other – have shown promise in tackling complex problems via deliberation and diverse perspectives [9][10]. These agents can then engage in multi-round dialogues, simulating a debate or collaborative design session. Early studies indicate that such **simulated deliberation** can improve reasoning and decision-making: for example, teams of LLM agents have been found to reduce individual model biases by exchanging different viewpoints, and to catch each other's errors or *hallucinations* through feedback-based dialogue [11][12]. Multi-agent discussions can also increase transparency, as the conversation between agents provides an explanatory trace of how a conclusion was reached [10][13]. Recent work by Zeng *et al.* [14] provides a concrete example: they developed *AutoDefense*, a multi-agent LLM framework that assigns separate roles to AI agents (such as content generator and policy judge) to collaboratively filter out harmful outputs from a model. This role-divided agent team was able to significantly improve the model's adherence to safety policies and resist malicious prompts. Such findings highlight the **potential of multi-agent LLM systems to achieve better ethical alignment** through internal deliberation and oversight, embedding a form of "ethical reflexivity" into AI-driven tools.

Another technological development informing this thesis is the technique of **Retrieval-Augmented Generation (RAG)** for LLMs. RAG is a method whereby an LLM is not only relying on its internal learned knowledge, but actively retrieves relevant external documents (such as databases, knowledge bases, or specific text corpora) to ground its outputs in up-to-date information [15]. This approach has been shown to improve the factual accuracy and transparency of LLM responses, since the model can cite retrieved sources and is less prone to hallucinating incorrect information [15]. In the context of AI ethics and software development, **live retrieval of ethical and legal documents** offers a promising way to keep an AI system's decisions aligned with established norms. This thesis is particularly motivated by the *European Union AI Act* and similar regulations, which will require software organizations to adhere to specific requirements for transparency, risk management, and human oversight in AI systems. We envision that a **RAG-empowered multi-agent system** could dynamically retrieve relevant portions of documents like the EU AI Act, ISO/IEC AI standards, or an organization's own code of ethics, and feed those into the agents' discussion. In doing so, the AI agents' deliberations are not just guided by generic moral principles but are **explicitly grounded in concrete rules and definitions** from authoritative sources.

In summary, the background to this thesis is characterized by a recognized shortfall in current software development practices with respect to AI ethics, and the emergence of new AI techniques that could address this gap. Traditional, principle-based ethics frameworks provide important high-level goals but insufficient operational guidance, leaving many teams unsure how to implement ethics day-to-day. This has motivated a search for more practical frameworks and tools. **Multi-agent LLM systems** with role specialization offer a way to embed ethical reasoning within AI, mimicking the checks and balances of human deliberation, while **RAG** offers a means to ground those deliberations in factual, up-to-date ethical standards and laws. These developments form the foundation and motivation for the present research. The central question becomes: *Can software development adopt AI ethics by embedding these AI advancements?* The next sections formalize this inquiry by outlining the research problem, questions, objectives, and approach of this thesis.

1.2 Research Questions

Given the background above, the **research problem** addressed by this thesis can be summarized as follows: Software engineering lacks effective mechanisms to ensure that development of AI systems are ethically aligned in practice, beyond mere adherence to abstract principles. This problem manifests as a disconnect between ethical intent and operational reality – developers and AI systems may unintentionally produce outcomes that violate ethical norms or legal requirements because current development processes do not systematically enforce those norms. The thesis posits that a solution framework with multiple interacting AI agents and real-time information retrieval, could help bridge this gap. We are investigating **whether a Multi-Agent RAG framework can assist software teams in adopting and enforcing AI ethics during development.** The core problem decomposes into two facets: first, understanding how specific design features of a multi-agent AI system contribute to ethical behavior; second, determining how such a system can be used as a tool to align AI outputs with external standards.

To address this overall problem, the research is guided by **two key research questions (RQs)**:

RQ1 (Literature Synthesis): *How do agent specialization, multi-round interactions, structured prompts, and conversational protocols contribute to the ethical alignment of outputs generated by a multi-agent LLM system?*

Explanation: This question focuses on the design parameters of multi-agent LLM systems. It seeks to synthesize insights from existing research (through a systematic literature review (SLR)) about the impact of a range of factors on ethical alignment. **Agent specialization** refers to giving agents distinct roles or expertise (e.g., one agent focused on risk analysis, another on ethical adherence of output), which might enforce a diversity of ethical perspectives. **Multi-round interactions** denote the iterative dialogue among agents – for instance, does allowing agents to debate or refine answers over several turns yield more ethically sound conclusions than a single-turn response? Likewise, **structured prompts and conversational protocols** involve the way the dialogue is orchestrated – e.g., turn-taking rules, prompt templates or inner monologue that ensure certain ethical checks (such as “Reply-Reflect-Critique-Code”). By answering RQ1, the thesis aims to distill *theoretical foundation* about these design choices. This will identify which mechanisms are most promising for steering a multi-agent system toward ethical outputs (for example, prior work suggests structured discussion formats can improve reasoning and reduce biased answers (Yin *et al.*, 2023; Yang *et al.*, 2024)).

RQ2 (Tool Development): *In what ways can a Multi-Agent RAG framework assist in aligning LLMs with established ethical standards and regulatory guidelines?*

Explanation: This question is more applied and integrative. Building on the understanding from RQ1, RQ2 asks how to design and implement a concrete **framework with tool**

– "Agents4EthicalSE" – that uses multiple LLM agents plus retrieval capabilities to ensure compliance with ethics and regulations in practice. RQ2 is about *operationalizing* the insights by creating an artifact (a software prototype) and evaluating its effectiveness. Key phrases here are *"established ethical standards and regulatory guidelines,"* which include documents like the EU AI Act, industry standards (e.g., ISO/IEC 24028 on AI trustworthiness), and digital ethics charters. The question implies multiple dimensions:

- **Framework design** – How should the multi-agent system be structured (what agents, what roles, what retrieval sources) to be useful for software developers concerned with ethics?
- **Assistance functionality** – In what ways can such a system assist? For example, can it act as an interactive *ethics consultant*, providing recommendations or warnings during development? Can it serve as a *compliance checker*, automatically flagging potential violations of laws?
- **Alignment efficacy –** Does incorporating a Multi-Agent RAG system measurably improve the ethical alignment of an AI model's outputs compared to not using such a system?

By addressing RQ2, the thesis will demonstrate the feasibility and benefits of the proposed approach. It goes beyond theory to show how a tool could integrate into software development and to observe its behavior. For instance, an envisioned scenario might be: a developer asks an AI pair-programming assistant to implement a feature involving user data; the Multi-Agent RAG system then intervenes by retrieving relevant privacy regulations and having an "Ethicist Agent" and a "Python Developer Agent" discuss the request, resulting in a recommendation that respects user consent requirements per the GDPR. In short, RQ2 guides the creation and evaluation of a tangible solution that embodies the findings of RQ1.

1.3 Objectives and Scope

Reflecting the research questions, the primary **objective** of this thesis is to explore and demonstrate an approach for integrating AI ethics into software development via a **Multi-Agent RAG framework**. This breaks down into three concrete objectives:

O1: Conduct systematic literature review (SLR) on ethical alignment mechanisms in multi-agent LLMs.

The thesis will conduct a structured **systematic literature review (SLR)** to gather and analyze prior research on multi-agent LLM systems, conversational AI alignment, and related areas. The goal is to identify how features like agent role specialization, iterative multi-turn dialogues, structured prompting strategies, and defined interaction protocols have been used (or could be used) to promote ethical outcomes. The outcome will be a clear understanding of which design strategies are reported to mitigate issues such as bias, toxic outputs, or rule violations in multi-agent settings (e.g., learning that having an agent explicitly tasked with rule-enforcement can prevent policy breaches (Zeng *et al.*, 2024).

O2: Design the framework and implement - Agents4EthicalSE

Using the insights from Objective 1, the thesis will design a **multi-agent LLM framework augmented with retrieval (RAG)**, tailored for the context of software engineering. A significant objective is to implement a prototype of this framework – effectively building a *tool* that can be used by developers in simulations or actual case studies. The design will follow established principles from AI and software architecture and will explicitly integrate ethical sources into the agents' decision loop via RAG. This objective addresses the "how" aspect of RQ2, by creating the means through which a multi-agent system can assist with alignment.

O3: Evaluate the framework's effectiveness in aligning AI outputs

A further objective is to **evaluate the Agents4EthicalSE tool** through empirical means – a workshop survey, to see in what ways and how well it assists in understanding of ethics by the developers and ethical alignment (the essence of RQ2). This involves evaluating the framework on scenarios or case studies derived from real AI Incidents Database that pose ethical dilemmas. For example, one evaluation scenario might be a content filtering task: given a user prompt that could lead a normal LLM to produce disallowed content, does the multi-agent framework successfully intervene and produce a response consistent with ethical guidelines (as in the spirit of AutoDefense by Zeng *et al.*)? Another scenario could involve compliance: checking if an AI component meets criteria of the EU AI Act's risk class definitions. The evaluation will be twofold – a *qualitative* assessment (examining the dialogues between agents to see if they reflect ethical reasoning and proper use of retrieved information) and a *quantitative* assessment (measuring outputs against benchmarks for harmful content, bias, or legal compliance). Achieving this objective will help determine the strengths and limitations of the proposed approach and provide lessons on its practicality in a real software engineering context.

The **scope** of this thesis (shown in Table [1]) is deliberately focused on the intersection of software engineering, AI ethics, and AI agent systems. The following clarifications on scope are important:

Table 1. *Scope of the Thesis*

Thesis Focus	Intersection of software engineering, AI-ethics, and AI agent systems
Primary Subject	Ethical alignment in software development using LLM-based (text-generating) AI systems
Modality Scope	Focus on text-generating AI (e.g., NLP models, coding assistants); other modalities (e.g., vision) mentioned only conceptually
Ethical Standards Used	European Union AI Act (EU AI Act), High-Level Expert Group on Artificial Intelligence (AI HLEG), Charter Of Fundamental Rights of the European Union, European Declaration on Digital Rights and Principles.
Normative Position	Framework does not create new ethical principles; it enforces and operationalizes existing standards
Ethical Issues Covered	Bias, non-discrimination, privacy, transparency, accountability (as defined in the above documents)
Methodology	Mixed approach: (1) Systematic Literature Review (SLR) for RQ1, (2) Design Science for RQ2
SLR Scope (RQ1)	Peer-reviewed studies for defined search string from databases like IEEE, ACM, Scopus
Design Science (RQ2)	Based on Hevner *et al.* [16] method: iterative development and evaluation of an artifact
Artifact Name	*Agents4EthicalSE*
Artifact Nature	Research prototype (proof-of-concept); not an industrial or productized system
Evaluation Environment	Controlled examples/simulations sufficient to demonstrate principles and answer RQs
Survey Participants	82 master's students from programme in Computer and Data Sciences
Disciplinary Breadth	Interdisciplinary: combining computer science (AI, SE) with ethics and regulatory studies
Philosophical Depth	Does not engage in deep philosophical ethics; uses pragmatic "ethics-by-design" and risk management approaches
Legal Theory Depth	Does not perform legal theory analysis; treats regulatory documents as fixed constraints
Accessibility Strategy	Technical concepts and ethical/regulatory material will be explained clearly for both technical and non-technical readers

By maintaining this focus and scope, the thesis ensures *how software development might adopt AI ethics through a multi-agent RAG approach*, without being sidetracked by issues beyond its domain. The deliverables will include the literature review findings, the designed framework and its documentation, and the results of the framework's evaluation in aligning AI outputs with ethical expectations.

1.4 Contribution of the Thesis

This research aspires to make primarily three contributions to both academia and practice at the intersection of AI ethics and software engineering:

C1: Synthesis of literature bridging multi-agent systems and AI ethics

While there is growing literature on multi-agent LLM collaboration and on AI ethics implementation tools, an integrated account is currently lacking. The thesis will contribute by mapping out the design space of **agent roles, interaction patterns, and prompting techniques** specifically in service of ethical AI. This contributes to knowledge by highlighting, for example, which structured multi-agent debate can yield more robust ethical decision-making (Yang *et al.*, 2024) or that incorporating an "ethics agent" role is a promising pattern to emulate. Such insights can inform future research and development of AI systems that are *"ethics-aware by design."*

C2: Empirical insights into multi-agent RAG efficacy for alignment:

Through the evaluation of RQ2, the thesis will offer empirical findings on how effective and in what ways the multi-agent RAG approach contributes to ethical alignment. These results will contribute evidence to the debate on AI alignment techniques. Notably, if successful, it will show that *role-playing agents plus retrieval can enforce complex normative constraints* (such as those in the EU AI Act) more reliably than a standalone model. It will also shed light on any limitations – for instance, the evaluation might reveal scenarios where the framework struggled (such as very ambiguous ethical situations or conflicts between principles).

C3: Interdisciplinary methodology for ethical AI in SE:

Indirectly, the thesis contributes a **methodological template** for tackling similar problems. By combining an SLR (from evidence-based software engineering) with a design science artifact development (common in information systems research), it shows how to rigorously go from identified theoretical gaps to a working solution, potentially aiding others who build upon or refine this framework.

C4: Alignment with regulatory and industry needs

As a timely contribution, this work aligns with immediate needs of industry and policymakers as AI regulations tighten. The EU AI Act and similar regulatory frameworks are on the horizon of enforcement (European Parliament, 2024). Companies will need practical tools to ensure their AI systems comply with these rules. This thesis contributes by preemptively exploring one such tool. The conceptual framework and results could inform **compliance solutions** in the software industry.

1.5 Structure of the Thesis

The remainder of this thesis is structured as follows:

- **Chapter 2 – Related Work:** This chapter provides an overview of the theoretical and practical foundations underpinning the thesis. It reviews relevant literature on AI ethics frameworks, ethical software engineering practices, and the limitations of current approaches (elaborating on topics introduced in the background).
- **Chapter 3 – Research Methodology:** This chapter describes the research design and methods used to answer RQ1 and RQ2. It outlines the protocol for the SLR, including how studies were selected, analyzed, and synthesized. It also details the design science research methodology adopted for developing and evaluating the Agents4EthicalSE tool (following guidance from design science literature such as Hevner *et al.*, 2004).
- **Chapter 4 – Literature Review Results (RQ1):** In this chapter, the findings from the literature review are presented, directly addressing RQ1. It organizes the results around the key factors in the question: agent specialization, multi-round interaction, structured prompting, and protocols. For each, it summarizes what the surveyed literature says about its influence on ethical or aligned outcomes.
- **Chapter 5 – Design of the Multi-Agent RAG Framework (Agents4EthicalSE):** This chapter covers the conception and implementation of the proposed framework, addressing RQ2 (the design aspect). It begins by specifying the requirements and design goals for the framework based on previous chapters (e.g., it should allow role specialization, iterative debate, and incorporate external knowledge sources). Then, it describes the system architecture – for instance, the different agent types included, the retrieval mechanism, and how the agents communicate (via a defined protocol). We will describe how we integrated sources like the EU AI Act into the retrieval index, and how prompts were structured to guide the agents' behavior (for example, the Ethicist agent is prompted with specific instructions to quote relevant law sections when disagreeing with a proposal). Implementation details, such as the LLM platform used (OpenAI GPT-4o, Assistant API), and how the agents were orchestrated (synchronously vs

asynchronously, etc.), are documented. A running example is provided to illustrate the tool's operation.

- **Chapter 6 – Evaluation and Discussion:** This chapter evaluates the Agents4EthicalSE framework in terms of its ability to assist ethical alignment during software development. The evaluation is conducted through mixed-method structured workshop survey of 82 master students. Participants were instructed to retrieve real AI incident cases from the AI Incident Database, formulate project descriptions, and assess responses generated by Agents4EthicalSE against a standard AI code generator. Metrics evaluated include ethical reasoning, coverage of legal concerns, output usefulness, and explainability. The survey responses and conversation artifacts formed the empirical basis for the assessment.
- **Chapter 7 – Threat to Validity:** This chapter reflects on the internal, external, construct, and conclusion validity of the research. It discusses potential biases in literature selection (SLR), limitations of the tool evaluation (e.g., user experience variation, limited number of prompts), and generalizability of findings. It also acknowledges technological limitations (e.g., reliance on specific LLM versions, retrieval latency), and suggests how these threats were mitigated or should be addressed in future work.
- **Chapter 8 – Concluding Remarks:** The concluding chapter revisits RQ1 and RQ2 and state how each was addressed. The chapter will highlight the contributions again and it will discuss the implications of this work for researchers. (e.g., new research directions or hypotheses arising from our findings) and for practitioners (e.g., how developers or organizations might begin to use AI agent assistants for ethics). Finally, the thesis concludes with suggestions for future work.

RELATED WORK

2.1 AI Ethics in Practice: From Principles to Engineering

AI ethics has become an important concern, leading to development of high-level frameworks and principles intended to guide AI development. By 2019, at least 84 institutional documents on AI ethics had been published, converging on common values such as transparency, fairness, privacy, accountability, and safety [1][2]. Examples include the EU's *Ethics Guidelines for Trustworthy AI* (2019), the OECD AI Principles (2019), and UNESCO's AI ethics recommendations (2021), all of which articulate similar broad tenets. This global intersection on principles reflects widespread agreement on *what* ethical AI should entail.

Critics argue that declaring ethical principles can even be counterproductive if not followed by concrete measures. In response, governments are moving toward strict requirements: for instance, the forthcoming EU AI Act will legally mandate risk management, transparency, and human oversight for certain AI systems. Yet, preparing software teams to comply with such regulations remains challenging in the absence of practical tools.

2.2 Operationalizing AI Ethics in Software Engineering

Recognizing the need to move *"from principles to practice,"* researchers and practitioners have proposed various methods to embed ethical considerations into the software development lifecycle. **Toolkits, checklists,** and **process guides** have emerged to help developers identify and mitigate ethical risks during requirements engineering, design, and testing. For example, Vakkuri *et al.* [16] introduced ECCOLA, a method integrated into agile sprints that provides developers with reflective questions and checklists to consider ethical issues at each stage of development. By injecting ethics prompts into routine sprint activities, ECCOLA attempts to make abstract principles more concrete and actionable for engineering teams. Similarly, Siqueira *et al.* developed the RE4AI Ethical Guide, a set of 26 "ethics requirement" cards aligned with 11 AI ethics principles (e.g., fairness, privacy, accountability) [17]. This guide is used during requirements elicitation to ensure teams proactively surface ethical requirements alongside functional requirements. These examples illustrate the broader trend of creating **practical checklists and frameworks** that translate universal ethics tenets (like transparency or justice) into domain-specific questions, criteria, or test cases for software projects [8].

Other operational tools focus on specific ethical dimensions. For instance, there are open-source toolkits for fairness auditing (e.g., IBM's AI Fairness 360) and model interpretability (e.g., LIME and SHAP). Documentation practices like *Datasheets for Datasets* and *Model Cards* have also been proposed to instill greater transparency [18]. Organizational process models—such as Microsoft's Responsible AI Standard or Google's AI Principles implementation guidelines—recommend governance structures like ethics review boards. Yet, despite these tools, uptake remains inconsistent [19]. Developers often report challenges of usability, relevance, and fit within agile workflows.

Preceding studies also highlight a lack of ethical training and unclear accountability structures as persistent barriers. AI code generation tools (e.g., GitHub Copilot) currently lack built-in ethical safeguards, further compounding the problem. Hence, even as ethical awareness rises, there remains a need for more seamless, embedded ethical guidance throughout the software development pipeline. These shortcomings motivate exploring innovative approaches, such as intelligent assistants that can *actively guide ethical software development in real time*.

2.3 Multi-Agent LLM Systems and Ethical Deliberation

One promising direction for aligning AI outputs with ethical standards is the use of **multi-agent LLM systems** – multiple language model "agents" interacting with each other to critique, debate, and refine their responses. Recent research suggests that having LLMs engage in *internal deliberation* can improve both the quality and alignment of their outputs. Yin *et al.* [9] proposed the Exchange-of-Thought (EoT) framework to structure cross-model communication. Their empirical results on reasoning tasks demonstrate that deliberation among diverse agents improves factual accuracy and consistency.

Chen *et al.* [10] introduced ReConcile, a round-table protocol in which diverse agents engage in consensus-building dialogues. This method improves reasoning by allowing agents to iteratively refine answers. Yang *et al.* [11] explored LLM voting mechanisms and showed that consensus among AI agents often yields human-aligned decisions but also risks over-conformity.

Multi-round interaction protocols further strengthen alignment through iteration. In a **debate framework**, two or more LLM agents exchange arguments over multiple turns, each critiquing or building upon the others' statements. This iterative turn-taking has been found to reduce hallucinations and improve reasoning correctness, as agents correct each other's mistakes. Du *et al.* [12] proposed a multi-agent debate method, showing that adversarial exchanges reduce hallucinations and improve truthfulness. Sun *et al.*

[13] extended this line of work with Corex, a collaborative framework that blends debate, review, and retrieval-based strategies. Together, these studies show that internal deliberation among specialized agents can enhance ethical reasoning. Similarly, diverse multi-agent debates have been proposed to encourage exploring alternative reasoning paths, breaking the "fixed mental set" of a single model's thinking [20].

2.4 Agent Specialization and Structured Prompting

A critical design factor in multi-agent LLM systems is how the behavior of each agent is constrained or guided – in other words, **prompt engineering and role specification**. Structured prompting involves crafting the inputs to each LLM (the "system prompts") such that the model adopts a clear persona or follows a defined format in its responses. Reynolds and McDonell [21] showed that prompting GPT-3 as an "AI safety analyst" led to more aligned responses. Similarly, Bai *et al.* [22] demonstrated an approach where an LLM was guided by an explicit *"constitution"* of principles during its response generation, resulting in outputs that avoided toxic or biased content without the need for explicit human feedback at each turn. Chen *et al.* [10] and Du *et al.* [12] implemented multi-round interaction protocols with defined roles and observed better judgment consistency. In these systems, agents cross-examine each other's responses using structured critiques.

Beyond roles, the *structure* of prompts and interactions is crucial. Techniques like **Chain-of-Thought prompting** and **Reflective Decoding** exemplify how structuring the reasoning process can yield better results. Chain-of-Thought prompting encourages an LLM to break down its solution into a step-by-step reasoning chain (often by instructing it to "think step by step"), which has been shown to reduce reasoning errors and increase transparency of the model's logic. Reflective Decoding goes a step further by having the model generate an initial answer, then critique its own answer, and finally produce a revised answer.

These findings suggest that structured prompts and protocol design play a vital role in facilitating multi-agent oversight. This is especially relevant in ethical decision-making, where distinct perspectives must be represented and negotiated.

2.5 Retrieval-Augmented Generation for Regulatory Alignment

As illustrated in Fig [1] RAG systems combine a generative model with a retrieval mechanism: the model can query a knowledge base (documents, databases, websites) and incorporate the retrieved information into its responses. This approach effectively gives the LLM an **open-book exam**, allowing it to draw on up-to-date and authoritative sources rather than relying solely on its static trained knowledge. Lewis *et al.* [15] showed that augmenting a model with a Wikipedia lookup significantly boosted performance on knowledge-intensive QA tasks improving factual accuracy. In the context of ethics, RAG systems can fetch **established ethical standards and regulatory guidelines** (e.g., EU AI Act) and inject them into prompts to guide compliant outputs.

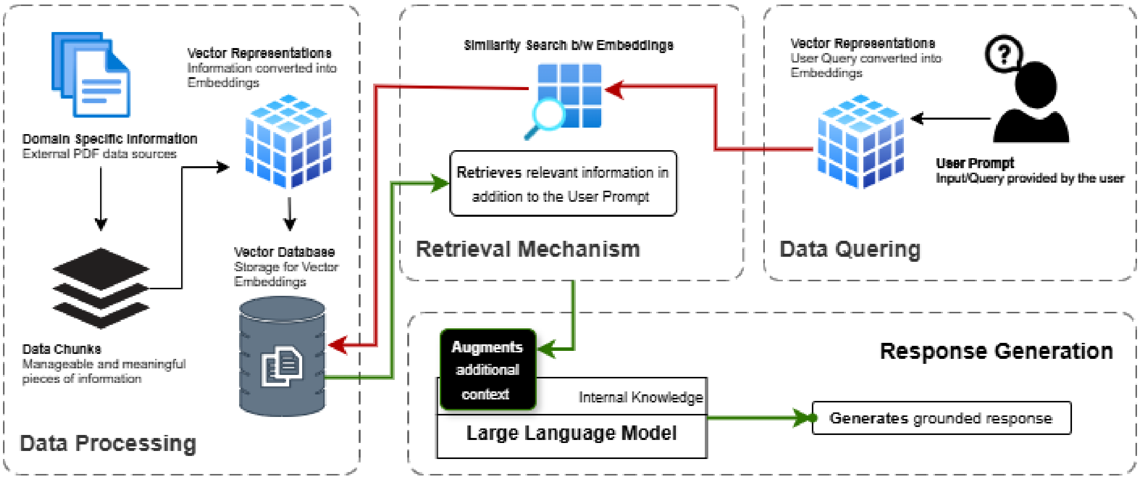

Figure 1. *Retrieval-Augmented Generation (RAG) Mechanism.*

For example, AutoDefense used a multi-agent framework to filter harmful content through role-specific LLMs—one for intent detection, another for policy judgment [14]. Sun *et al.* [13] employed retrieval to reinforce agent deliberation. These efforts reveal the potential of combining RAG with multi-agent systems to ensure ethical alignment and regulatory adherence.

2.6 Summary and Implications for This Thesis

Across the literature, we observe key trends: the limitations of principle-based ethics, the partial success of operational checklists, and the promise of multi-agent deliberation. Yet, most prior work treats RAG, prompt structuring, and agent specialization as isolated strategies. This thesis brings these threads together by proposing a Multi-Agent RAG framework — **Agents4EthicalSE** — to align LLM behavior with both ethical and regulatory standards.

METHODOLOGY

3.1 Research Design Overview

This thesis adopts a mixed-methods approach to address its two core research questions. The first question (RQ1) investigates how specific design elements—such as agent specialization, multi-round interactions, structured prompts, and conversational protocols—contribute to the ethical alignment of outputs generated by multi-agent LLM systems. To address this, a **Systematic Literature Review (SLR)** was conducted, following established guidelines for evidence-based software engineering [23]. The second question (RQ2) explores the feasibility and effectiveness of using a multi-agent, retrieval-augmented generation (RAG) framework to assist in aligning LLM-generated outputs with existing ethical standards and regulatory requirements. For this purpose, a **design science methodology** as highlighted in Fig [2] was employed [24], guiding the development and implementation of the proposed framework—Agents4EthicalSE.

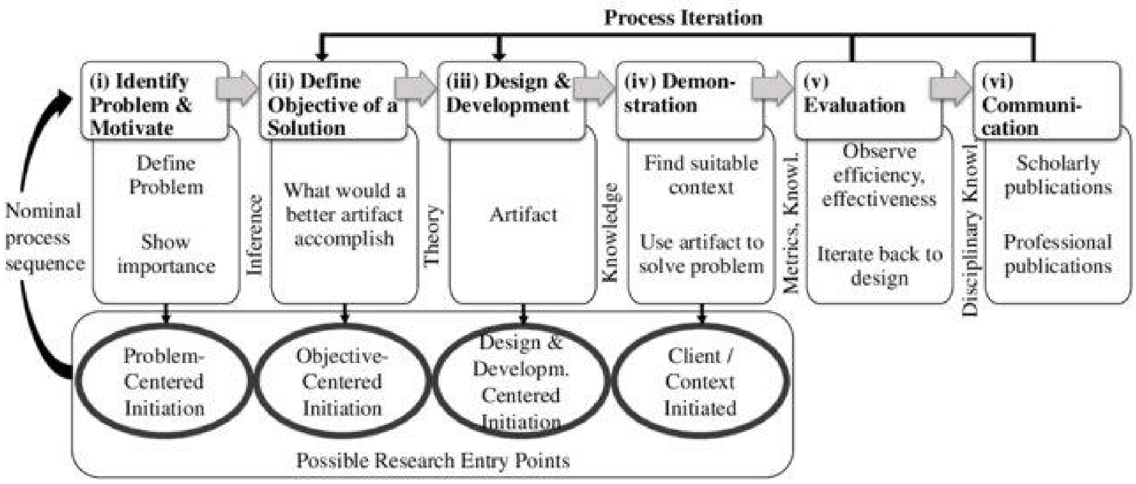

Figure 2. *Design Science Research Methodology Process Model. Adapted from Peffers et al. [25].*

By combining a rigorous synthesis of existing research with the construction of a functional prototype, the methodology aims to bridge theoretical understanding with practical contribution. The SLR component establishes a grounded understanding of what constitutes ethical alignment in multi-agent LLM systems, while the design component demonstrates how those insights can be operationalized through an AI-powered tool for ethical software development. This dual-track approach is consistent with recent research advocating for the integration of knowledge synthesis and artifact creation in AI ethics and responsible AI tool design [6][19].

The following sections describe in detail the methodological procedures followed in both the literature review and the design of the multi-agent RAG framework.

3.2 RQ1 - Systematic Literature Review (SLR)

3.2.1 Purpose and Scope

The primary objective of the systematic literature review (SLR) was to explore how design mechanisms within multi-agent LLM systems contribute to the ethical alignment of AI-generated outputs. This review focuses on four specific dimensions identified in contemporary AI ethics and multi-agent system research:

1. agent role specialization,
2. multi-round or iterative interaction,
3. structured prompting and instruction engineering, and
4. the use of conversational protocols for dialogue governance.

These aspects are hypothesized to influence how well such systems align with ethical principles, regulatory standards, or organizational guidelines.

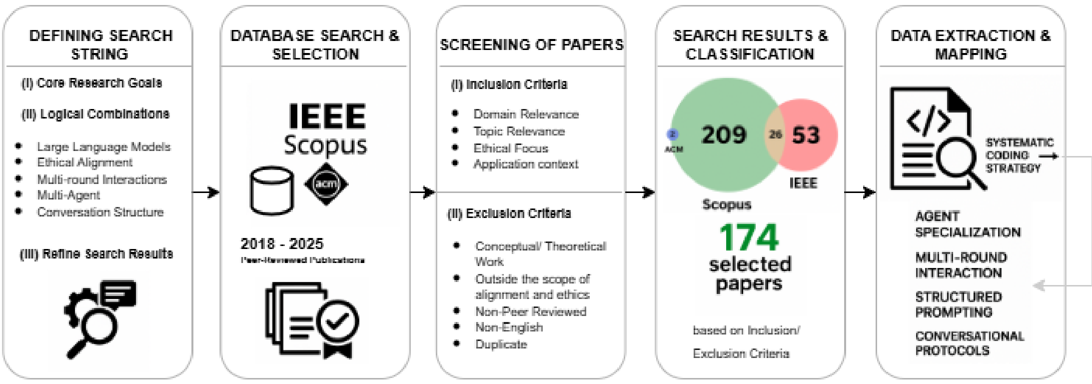

Figure 3. *Systematic Literature Review (SLR) Process*

The findings from the SLR process as depicted in Fig [3] informed the subsequent design choices for the Agents4EthicalSE framework, ensuring that the proposed solution is anchored in established design patterns and empirical evidence.

3.2.2 Review Protocol

The review protocol followed the procedures outlined by Kitchenham and Charters [1], ensuring transparency, repeatability, and rigor throughout the systematic review process. The protocol consisted of five core components: inclusion and exclusion criteria, selection of databases, formulation of search strings, definition of the period, and screening and quality assessment as shown in Table [2]. Each article was independently reviewed using these criteria and documented in an SLR log sheet *(see Appendix A for the access link [1]).*

Table 2. *Overview of Review Protocol for SLR*

1. Goal		
RQ1	How do agent specialization, multi-round interactions, structured prompts, and conversational protocols contribute to the ethical alignment of outputs generated by a multi-agent LLM system?	
2. Protocol		
2.1 identification of inclusion/exclusion criteria	2.1.1 Inclusions Criteria	
	Time frame	Publications from 2018 onward (post-LLM emergence era, e.g., BERT, GPT-2 onwards).
	Language	English-language publications.
	Document type	Peer-reviewed journal articles, conference papers, systematic reviews.
	Domain relevance	Focus on multi-agent systems, LLMs, agent-based architecture, prompt engineering, or ethics in AI.
	Topical relevance	Studies must discuss agent specialization, multi-round/iterative interactions, structured prompts, or conversational protocols in relation to LLMs or generative AI.
	Ethical focus	Must address ethical alignment, value alignment, responsible AI, bias mitigation, or trustworthiness of model outputs.
	Application context	It can include theoretical, architectural, experimental, or use-case implementations.
	2.2.2 Exclusion Criteria	
	Non-LLM focus	Studies on classical multi-agent systems (e.g., swarm robotics, game-theoretic MAS) without LLM or generative AI relevance.

	Non-ethical focus	Papers only address performance, latency, or optimization without mention of ethical implications.
	No agent interaction	Studies that do not involve agent collaboration, dialogue, or prompt design elements.
	Gray literature	Blogs, informal essays, white papers without academic rigor (unless they are highly cited preprints).
	Low quality/short papers/non-peer-reviewed papers	Abstract-only papers, workshop summaries without full methodology or evaluation, roadmaps, vision paper.
	Language	Non-English publications.
2.2 selection of databases	**IEEE**	
	ACM Digital Library (ACM DL)	
	Scopus	
2.3 formulation of search strings	***("LLM" OR "large language model" OR "foundation model" OR "generative AI" OR "gen-AI") AND ("ethic*" OR "responsible AI" OR "AI alignment" OR "value alignment" OR "goal alignment") AND ("multi-agent*" OR "agent-based*" OR "AI agent" OR "collaborative agent" OR "agent specialization" OR "specialized agent" OR "role-based agent" OR "agent role" OR "task-specific agent" OR "multi-round interaction" OR "multi-turn interaction" OR "iterative prompt" OR "agent dialogue" OR "agent coordination" OR "structured prompt" OR "prompt engineering" OR "system prompt" OR "instruction prompt" OR "conversational protocol" OR "communication protocol" OR "interaction design" OR "dialogue framework")***	
2.4 definition of the time frame	**2018 - 2025**	
2.5 screening and quality assessment		

3.2.3 Data Extraction and Synthesis

To enable meaningful mapping between the literature and the 4 design dimensions specified in RQ1, a systematic coding strategy was employed. Each included study was manually reviewed and coded against a set of keywords and conceptual indicators representing the 4 dimensions: agent specialization, multi-round interaction, structured prompting, and conversational protocols. This was done using a qualitative content analysis method, where descriptive codes were derived from both theory and recurring language patterns found in the abstracts, author keywords, inclusion rationales, and—where necessary—the full methodology sections of the papers. While abstracts and metadata provided initial coding cues, in all cases where classification was ambiguous or nuanced, the methods and results sections were consulted to ensure accurate dimension mapping. This helped guard against misclassification and strengthened the validity of the analysis, especially for features like agent interaction protocols or role assignment mechanisms that may not be fully described in abstracts.

- For **agent specialization**, indicators included explicit reference to agents being assigned unique roles or personas (e.g., “safety monitor,” “planner,” “devil’s advocate”), role-based decision structures, or division of responsibilities among agents.
- For **multi-round interactions**, papers were tagged if they described iterative decision-making, multiple turns of interaction between agents, or feedback refinement protocols.
- For **structured prompts**, coding focused on whether prompt engineering, template design, or instruction tuning was central to the method.
- For **conversational protocols,** relevant indicators included the use of rule-based turn-taking, dialogue control strategies, meta-conversational norms, or system-governed agent exchanges.

The coding was operationalized against the four design categories with supporting text fragments. These mappings were then validated using inter-keyword agreement heuristics and frequency filtering. The frequency counts of each mapped dimension helped quantify research density across the four design areas and formed the basis for the thematic synthesis that follows in the next section.

Once eligible studies were selected, relevant information was extracted to map each paper’s contribution to one or more of the four core design features: agent specialization, multi-round interactions, structured prompting, and conversational protocols *(see Appendix A for the access link [2]).*

For each article, metadata such as publication venue, year, study type (empirical, conceptual, technical), and model used (e.g., GPT-3, Llama) were recorded. In addition, qualitative data were extracted regarding how ethical alignment was conceptualized and measured, the role of agent interaction in producing ethical outputs, and the specific prompt engineering techniques employed.

A thematic synthesis approach was used to analyze the data, allowing concepts and strategies from individual studies to be grouped into emergent themes. Thematic synthesis is especially suitable for reviews that aim to interpret diverse forms of knowledge (e.g., design proposals, empirical results, conceptual frameworks) and to draw connections between them [26]. The themes were iteratively refined and categorized into higher-order codes representing patterns of agent collaboration and ethical guidance mechanisms. This synthesis facilitated the identification of common design patterns, trade-offs, and unresolved tensions in literature.

The results of this analysis directly inform the architecture and features of the proposed Agents4EthicalSE framework described in section 3.3.

3.2.4 Results

From a total of **174 peer-reviewed** studies after the screening process (as shown in Fig [4]) that began with an initial pool of 323 articles identified across three academic databases: IEEE Xplore, ACM Digital Library, and Scopus, **four major themes** directly aligned with the design features outlined in RQ1. These dimensions were supported by a diverse set of empirical and conceptual works.

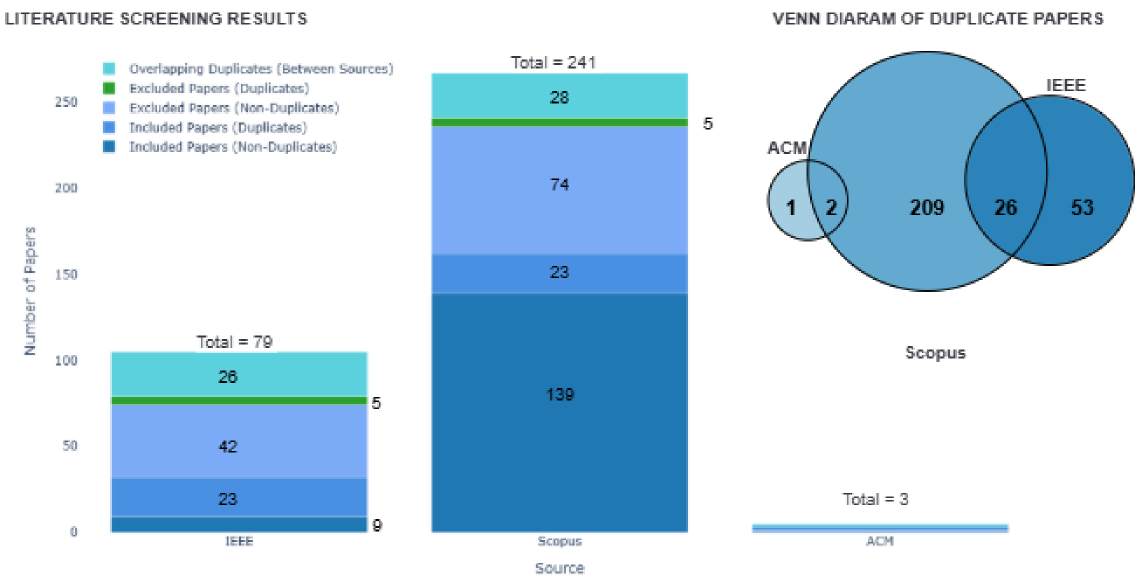

Figure 4. *Literature Screening Results from Databases*

- **Conversational Protocols (165 studies):** All selected studies incorporated some form of protocol-guided interaction among LLM agents or between agents and users. For instance, Yin et al. [27] proposed MDD-5k, a dataset synthesized through dialogue-driven neuro-symbolic agents, while Silva et al. [28] and Bandi et al. [29] introduced conversational governance layers to enforce ethical compliance in therapeutic and healthcare applications. Two studies also addressed adversarial robustness via protocol coordination (Chao et al. [30]) and examined system transparency through modular conversational agents (Yu et al. [31]).

- **Structured Prompts (124 studies):** A large body of work analyzed prompt engineering as a technique for steering LLM behavior. For example, Lynch et al. [32] evaluated how prompt templates affect sentiment in narrative generation, while Lim and Schmälzle [33] used empirical studies to assess prompt impact on message quality in health contexts. Frameworks like Reflexive Prompting (Djeffal [34]) and domain-specific designs (Ariely et al. [35]) show how instruction structure can constrain LLM responses toward normative objectives. These approaches validate the importance of structured prompting for embedding ethical safeguards in LLM workflows.
- **Agent Specialization (23 studies):** While less frequent, impactful studies emphasized the benefits of assigning differentiated roles to agents. Weng et al. [36] studied conformity behavior across specialized agents, while Xu et al. [37] developed a multi-objective framework for mitigating social bias via task-specific roles. Vaccaro Jr. et al. [38] explored agentic personalization of STEM texts using designated tutoring and editing personas. The collected findings suggest that role heterogeneity enhances deliberative quality and traceability.
- **Multi-Round Interactions (19 studies):** Studies in this category illustrated the iterative enhancement of reasoning through feedback and dialogue cycles. For instance, Sakaguchi et al. [39] compared human and AI-based thematic analysis over multiple refinement rounds, while Liu et al. [40] demonstrated how LLM debates and feedback mechanisms reduce hallucinations and improve factual consistency. These works emulate human review boards and underscore how temporal sequencing of interaction contributes to alignment.

Collectively, these results indicate that conversational and prompt-based mechanisms are now standard practice in alignment research, while agent specialization and iterative coordination remain fertile ground for innovation.

3.3 RQ2 - Design and Development of the Multi-Agent RAG Framework: *Agents4EthicalSE*

3.3.1 Design Goals and Rationale

The core purpose of the *Agents4EthicalSE* framework is to support software developers in aligning their code artifacts with AI ethics principles and regulatory standards, particularly the **EU AI Act**, by embedding structured, document-aware guidance into LLM-assisted development workflows. Unlike generic AI coding assistants (e.g., GitHub Copilot, ChatGPT), which offer productivity gains but lack regulatory grounding, *Agents4EthicalSE* operationalizes ethical guidance by combining **multi-agent coordination**, **retrieval-augmented generation (RAG)**, and **structured conversational protocols**.

This design is grounded in prior research emphasizing the value of **agent specialization** [36], **document-grounded generation** [15], and **conversational protocols** [28] for aligning LLM outputs with human values. For example, Zeng et al. [36] and Vaccaro et al. [38] demonstrate that assigning domain-specific roles to LLM agents leads to more robust outcomes in tutoring and system design tasks. Meanwhile, Lewis et al. [15] highlights how RAG significantly improves factual grounding in generative models. Furthermore, Morley et al. [4] stress the need to operationalize AI ethics principles through practical toolkits that bridge abstract guidelines and software artifacts.

Its design is also informed by prompt engineering strategies shown to increase ethical sensitivity and controllability of LLM outputs (Lynch et al. [32]; Djeffal [34]; Goloujeh et al. [41]) and by research emphasizing the impact of multi-turn, role-specific dialogues for reflective and responsible generation (Silva et al. [28]; Bearman & Ajjawi [42]; Liu et al. [40]).

3.3.2 Framework Architecture

This section outlines the key components, explaining their roles in facilitating ethical AI development. Each component of *Agents4EthicalSE* framework as shown in Fig [5], from the initial risk assessment to the conversation structure and iterative rounds, is designed as an alignment technique to ground the LLM response with current ethical standards and enhance trust in AI driven outputs.

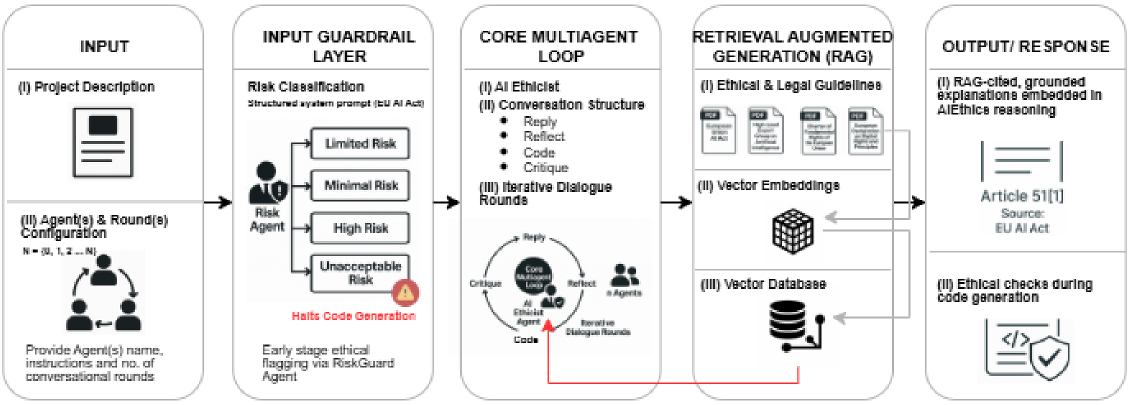

Figure 5. *Theoretical Framework for Multi-Agent RAG Framework.*

1) Input Guardrail Layer:

The framework includes an **Input Guardrail Layer** as the first stage of the process, where a dedicated RiskGuard Agent categorizes the project's ethical risk based on the EU AI Act's 4 risk levels for AI systems: Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk as shown in Fig [6]. This figure presents examples of AI module descriptions provided to the system, along with the RiskGuard Agent's response by categorizing each module under a specific risk level and providing justifications for its classification.

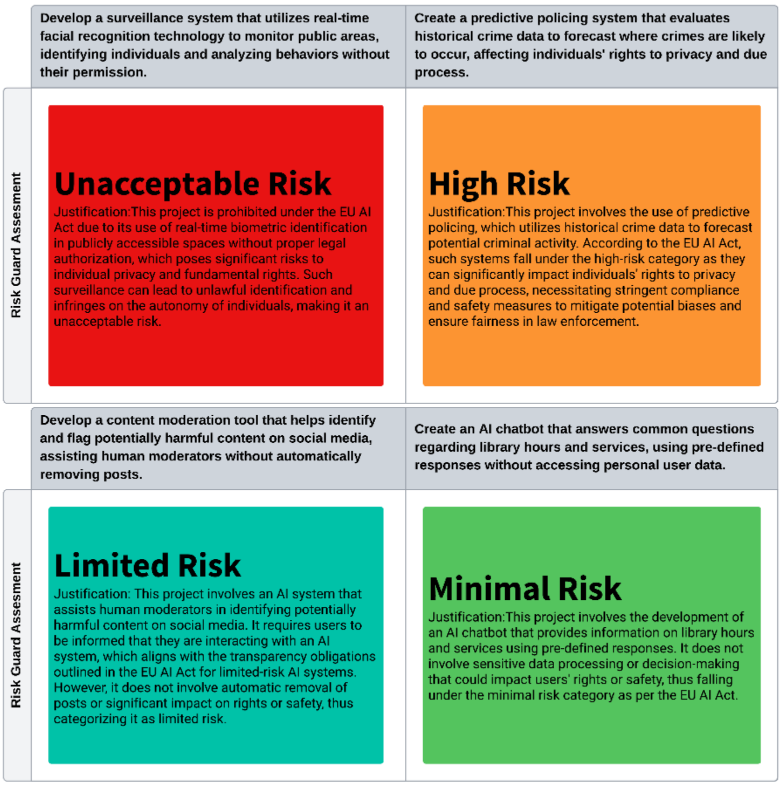

Figure 6. *Risk Level Assessment by RiskGuard Agent*

This initial layer acts as a safeguard to ensure that modules with a high or unacceptable risk level are identified early. If a module is evaluated as *"Unacceptable Risk",* the system does not provide code for implementation and prompts the user for a detailed ethical review by *AI Ethicist*. For modules classified under other risk levels, the process proceeds as usual. This approach reflects recent research advocating for early-stage risk assessment in AI systems [8][16], which helps developers address ethical concerns proactively, minimizing the likelihood of ethical and regulatory conflicts later in the development cycle.

2) Agent Roles and Specialization:

The framework includes pre-defined and user-extensible agents with specialized roles, each guided by a tailored prompt as highlighted in Program [1] and function:

- **RiskGuard** classifies ethical risk based on AI risk classifications defined by the EU AI Act.
- **AI Ethicist** A key component in the Agents4EthicalSE framework is the AI Ethicist agent, who serves as the final checkpoint in the conversation. This agent has access to external resources, including the EU AI Act, AI HLEG guidelines, and other relevant regulatory documents through the RAG system. The AI Ethicist’s role is to assess the ethical implications of each output, ensuring alignment with legal standards and addressing specific ethical principles, such as non-discrimination and fairness.

Additional agents, such as python developers, domain validators or accessibility reviewers, can be created based on user needs. Role specialization ensures epistemic diversity, a feature shown to improve ethical coverage and depth in multi-agent reasoning tasks [36][37].

```
ASSISTANT_INSTRUCTIONS = """
You are an AI risk assessment agent tasked with evaluating AI sys-
tems for compliance with the European Union's AI Act. Your primary
responsibility is to determine whether a given module description for
an AI system falls under the **"Unacceptable Risk"**, **"High Risk"**,
**"Limited Risk"**, or **"Minimal Risk"** categories based on its
specifications and potential impact on individuals and society.

### Step 1: **Review the module specifications** and identify the
**primary purpose, intended use, and target users**.
### Step 2: **Determine the context** in which the system will
operate (e.g., law enforcement, education, employment, public spaces).
### Step 3: Evaluate if the system meets **any** of the following
criteria:
{}
{}
{}
{}
```

```
    ### Output Format:
    Only respond using the following structured format:

    {{ "Category": "<Insert one of the following categories: 'Unac-
ceptable Risk', 'High Risk', 'Limited Risk', 'Minimal Risk'>", "Jus-
tification": "<Provide a justification for the selected category, ref-
erencing the European Union's AI Act and explaining the relevant cri-
teria>" }}

    #### Allowed Categories:
    1. **"Unacceptable Risk"**: This project is prohibited under the EU
AI Act due to its potential to cause significant harm or infringe on
fundamental rights.
    2. **"High Risk"**: This project must comply with stringent safety
and compliance requirements under the EU AI Act.
    3. **"Limited Risk"**: This project has specific transparency ob-
ligations but does not require stringent compliance measures.
    4. **"Minimal Risk"**: This project does not require specific com-
pliance measures under the EU AI Act.

    ### Instructions:
    - Select only **one** category from the list above.
    - Fill the `"Category"` field with the exact category name (e.g.,
`"High Risk"`).
    - In the `"Justification"` field, provide a concise justification
for your decision. Reference the relevant sections of the **EU AI
Act** to support your assessment.
    - Use formal language and ensure the output adheres to the specified
structure.

    ### Example Output:

    {{ "Category": "High Risk", "Justification": "This project involves
the use of biometric surveillance, which is classified under the EU
AI Act as a high-risk application requiring strict compliance and
safety measures to protect fundamental rights." }}

    """.format(UNACCEPTABLE_RISK_CRITERIA, HIGH_RISK_CRITERIA, LIM-
ITED_RISK_CRITERIA, MINIMAL_RISK_CRITERIA)
```

Program 1. *System Prompt Template provided to RiskGuard Agent.*

3) Dialogue Protocol and Multi-Round Interaction

The conversation follows a fixed four-phase sequence:

1. **Reply** – Agents respond to a project description or code snippet.
2. **Reflect** – They critique each other's input from their disciplinary lens.
3. **Code** – The agents update code with feedback incorporated.
4. **Critique** – The AI Ethicist evaluate ethical and legal compliance.

Table [3] outlines this protocol clearly. This dialogue structure draws from literature on **chain-of-thought prompting** [21], **reflective prompting** [35], and **deliberative scaffolding** in multi-agent LLMs [29][43]. Multi-round conversations allow for ethical reasoning to unfold over iterations, supporting value alignment and reducing hallucinations.

Table 3. *Conversation Structure of Agents*

Phase	Description	For Example: Python Developer	For Example: AI Ethicists
Reply	In this initial phase, each agent responds to the contributions of other agents, addressing relevant points and clarifying any aspects that align with their particular expertise. This ensures that all insights are contextualized within the specific agent's domain.	Coding best practices for a python developer.	Ethical considerations for the AI Ethicist.
Reflection	Agents then reflect on the conversation considering their role. This phase allows agents to incorporate their specialized knowledge into the development process actively.	A developer might consider the technical feasibility and optimization of the proposed code.	The AI Ethicist evaluates ethical implications based on regulatory guidelines and ethical principles.
Code	If required, the agents provide code examples that integrate their reflections into the development process.	For technical agents like the Python developer, this phase might involve drafting or optimizing code to meet quality standards.	For the AI Ethicist, this phase could involve suggesting modifications to the code that enhance compliance with ethical standards, ensuring that the system maintains a balance between functionality and ethics.
Critique	In the final phase, each agent offers constructive feedback, focusing on the aspects most relevant to their expertise. This step encourages agents to highlight potential areas for improvement, fostering a culture of iterative refinement	The Python developer could address coding inefficiencies or suggest improvements.	The AI Ethicist might point out ethical risks or compliance issues.

	and mutual accountability.		

4) Retrieval Augmented Generation (RAG)

The RAG mechanism enables the AI Ethicist to retrieve context-specific excerpts from a corpus of regulation documents inlcuding European Union AI Act, High-Level Expert Group on Artificial Intelligence (AI HLEG), Charter Of Fundamental Rights Of The European Union, and European Declaration on Digital Rights and Principles.

This rapid search capability is invaluable for practitioners, allowing the framework to efficiently ground responses and code recommendations in current standards, a task that would otherwise require extensive time and resources. For example, in a scenario evaluating a hiring AI module, the AI Ethicist cites risk-mitigation and transparency clauses directly (see *section 3.3.4*). This grounding boosts developer confidence and regulatory traceability.

3.3.3 Technology Stack and Implementation

Code Available: *(see Appendix B).*

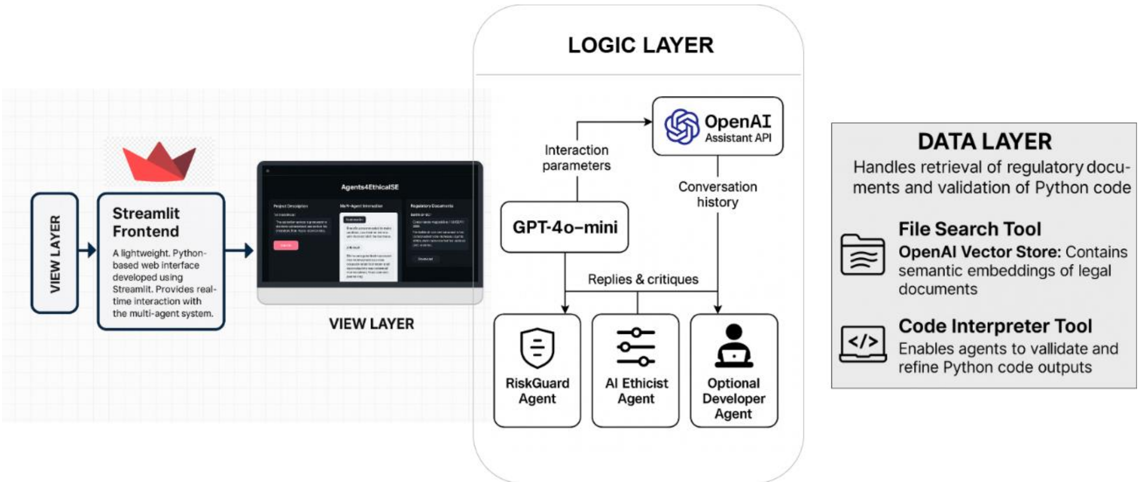

Figure 7. *System Architecture of Agents4EthicalSE Tool.*

The system implementation (Fig [7]) uses **OpenAI's Assistant API**, which powers the multi-agent environment with ***GPT-4o-mini*** as the underlying LLM for each agent. This setup provides a flexible environment where each agent has access to the tools and resources necessary to perform specialized tasks and ensures the system produces both technically sound and ethically aligned outputs. For example, each agent in the system has access to the ***OpenAI's Code Interpreter tool*** for Python, allowing it to handle code-based tasks efficiently. The AI Ethicist agent is also equipped with ***OpenAI's File Search*** tool for RAG, enabling it to access a vector store containing key ethical guidelines and regulatory documents. Integrating retrieval capabilities allowed the generative model to produce contextually relevant, grounded responses, thus enhancing accuracy and reducing the risk of outdated information.

1) Streamlit Frontend:

A lightweight, Python-based web interface was developed using Streamlit (Fig [8]). It allows users to interact with the multi-agent system in real time—submitting project descriptions, adding/deleting user agents, reviewing agents' dialogues, inspecting retrieved regulatory documents in a transparent and accessible layout and downloading the output as a conversation text file. *(see Appendix A for the access link [3]).*

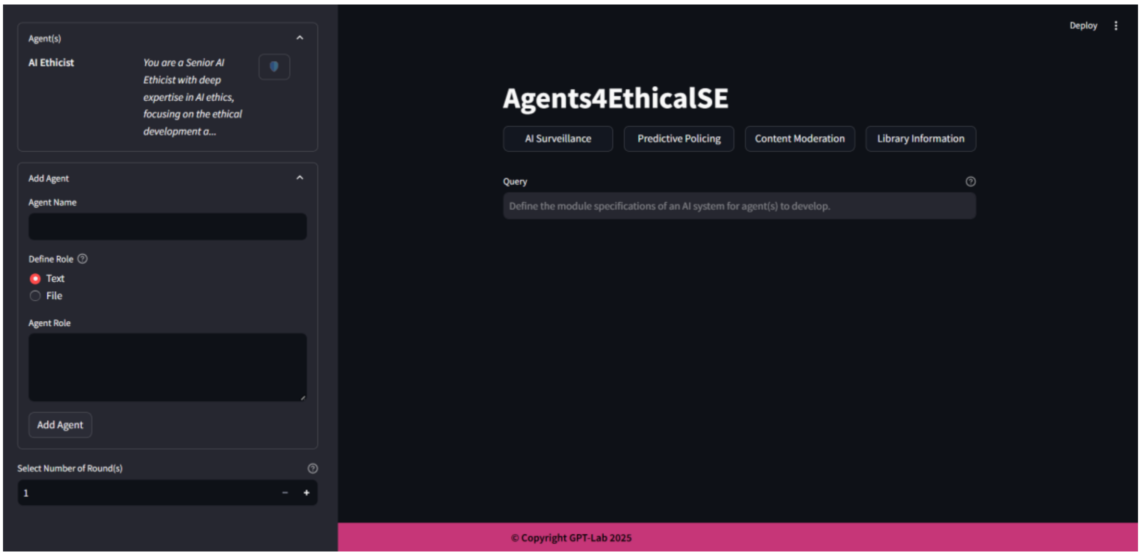

Figure 8. *Streamlit User Interface for Agents4EthicalSE.*

2) Deployment Platform:

The full application stack was containerized and deployed using a cloud platform (e.g.,CSC Server). This ensured remote access and ease of integration into distributed development workflows.

3) Python Environment:

All components were implemented in Python, selected for its compatibility with LLM orchestration frameworks, vector search libraries, and the broader AI tooling ecosystem.

4) OpenAI's Assistant API

serves as the core environment for creating and managing agents with specialized roles. Each agent is configured with a structured conversation flow and specific instructions in a *conversation thread*. This setup allows for clear communication among agents, as well as *persistent context* through the conversation history.

5) GPT-4o-mini

serves as the language model, chosen for its balance between performance and efficiency. It supports the interpretive and generative needs of the agents, especially for tasks that involve complex ethical assessments and multi-stage processing.

6) File Search Tool

OpenAI's Vector Store is populated with essential regulatory documents. OpenAI automatically parses and chunks documents, generates, and stores the embeddings, and uses both vector and keyword search to retrieve relevant content to answer user queries. The file search tool then enables the AI Ethicist to pull information directly from these sources, grounding its responses in specific, up to date references.

7) Code Interpreter Tool

enables Agents to write and execute *Python code* within a secure, sand-boxed environment. This tool can manage files with varied data formats and produce outputs such as data tables and visualizations, including graphs. With access to the Code Interpreter, Agents can enhance the accuracy and adaptability of code generation and refine code iteratively.

8) Agents Development and Specialized Prompts:

In implementing the agents within the Agents4EthicalSE tool, prompt engineering practices were systematically applied to enhance the reliability and alignment of each agent's output. These practices, as outlined in OpenAI's prompt engineering guidelines, emphasize clarity, structured instructions, and grounding responses in relevant reference materials. Each agent was developed with a well-defined role, clear instructions, and tailored prompts that guide them in completing specialized tasks, thereby minimizing ambiguity and improving response alignment with ethical standards.

• **RiskGuard Agent** For example, the prompt (*see Program 1*) provided to the RiskGuard Agent begins by clearly defining the role of the agent and then specifies the primary responsibility of the agent. It offers the agent a structured approach, ensuring it knows exactly how to proceed.

Tactic 1: The prompt specifies three steps required to complete the task, guiding the agent to focus on specific aspects (e.g., purpose, context, and criteria).

Tactic 2: It includes details in query, such as identifying the purpose and intended use, to get more relevant answers.

Tactic 3: It instructs the model to answer using EU AI Act defined risk categories as a reference text.

Tactic 4: It uses intent classification by dividing the risk levels into Unacceptable, High, Limited, and Minimal to identify the most relevant instructions. This strategy reduces cognitive load and error rates by focusing the agent's attention on predefined categories.

Tactic 5: The prompt specifies an output format with fields for Category and Justification, guiding the agent to produce concise, structured responses.

Tactic 6: The prompt's step-by-step approach gives the agent time to Think to evaluate the module's risk level systematically.

Tactic 7: The prompt includes an example output, which demonstrates the expected format and tone, helping the agent understand how to structure its responses.

• **AI Ethicist Agent** has access to File Search Tool. Each response/code generated by the AI Ethicist is enriched with context from the retrieved documents, helping ensure compliance with ethical principles. This practice also infuses the prompt engineering tactic to instruct the model to answer with citations from a reference text. Additionally, the AI Ethicist is instructed only to generate code when the module description does not fall under *Unacceptable Risk Category* with structured output format provided to guide factual reasoning of responses.

9) Parameter Adjustments for Controlled Output

Parameters such as *temperature* (randomness) and *top p*(diversity) are configured for each agent respectively, helping to manage the variability in agent responses. These settings support more controlled outputs that balance creativity with consistency, as recommended by OpenAI. For example, the value of ***temperature and top p is set to 0 and 0.5*** for RiskGuard and to moderate levels for AI Ethicist as it is also facilitating ethical code generation.

10) Contextual Continuity and Structured Conversation

The conversation history is retained throughout each round and provided as additional context to the next agent in loop. At the end of the round, conversation summary is shared with the first agent to ensure continuity. This setup follows a structured conversation flow that includes four phases: Reply, Reflection, Code, and Critique, injected to agent(s) instructions provided by the user. This approach fosters an iterative process with chain-of-thoughts in which agents refine their outputs, ensuring they remain aligned with the framework's initial context, guidelines and ethical standards.

This technology stack was chosen to match findings in existing literature on agentic systems and RAG-based reasoning. This modular architecture allows iterative refinement of output, promoting trust, fairness, and regulatory compliance in LLM-assisted software development.

3.3.4 Example Scenario and Output Walkthrough

The system shows in Fig [10] authoritative sources RAG searches through such as the European Union’s AI Act and the AI HLEG Ethics Guidelines, enabling the AI Ethicist to identify specific regulatory sections relevant to the *risk assessment of a surveillance system monitoring facial recognition in public spaces*. Fig [9] illustrates how the RAG component cites specific articles of the act with verified, context specific information, ensuring that decisions meet established standards of fairness, accountability, and transparency.

1. Risk Assessment

The evaluation of the surveillance system, which leverages real-time facial recognition technology to monitor public spaces without obtaining permission from individuals, has been deemed as posing an **unacceptable risk**. This classification aligns with stipulations found in the EU AI Act, specifically Sections concerning the use of **real-time biometric identification systems in publicly accessible spaces**.

The Act outlines that such systems should only be employed under stringent conditions with explicit legal authorization, aimed at pursuing specific objectives like the identification of missing persons or prevention of immediate threats to life or physical safety. The critical sections that substantiate this assessment include:

- **Article 5** specifically prohibits the placing, putting into service, or use of AI systems that utilize biometric identification in public spaces unless they are strictly necessary for limited and well-defined purposes[0][1].
- The requirement for a **fundamental rights impact assessment**, along with registration in a designated EU database prior to usage, underscores the need for stringent scrutiny of such systems[2][3].

In this instance, the lack of proper legal authorization and the absence of safeguards fundamentally contravene the norms established by the EU AI Act. Therefore, the project may not only infringe upon the privacy rights of individuals but also undermine essential democratic principles.

Figure 9. *Example of a Grounded Response for an Unacceptable Risk Level AI System with Cited Ethical Standards*

```
Source:
[
  0 : "[0] European Union's AI Act.pdf"
  1 : "[1] European Union's AI Act.pdf"
  2 : "[2] European Union's AI Act.pdf"
  3 : "[3] European Union's AI Act.pdf"
  4 : "[4] AI HLEG Ethics Guidelines for Trustworthy AI.pdf"
  5 : "[5] AI HLEG Ethics Guidelines for Trustworthy AI.pdf"
  6 : "[6] European Union's AI Act.pdf"
  7 : "[7] European Union's AI Act.pdf"
]
```


Figure 10. *Source List for RAG Retrieval*

In Fig [11], the AI Ethicist generates enhanced code for the module description, "*Develop an AI system that screens and ranks job applicants based on their resumes and video interviews to recommend top candidates for hiring.*" The initial code includes core functionality for processing resumes and ranking candidates. Through iterative feedback, as illustrated in Fig [12], the AI Ethicist identifies areas for ethical improvement, such as implementing bias mitigation, data governance, and explanation mechanisms—critical elements for transparency and fairness in AI-driven hiring processes.

3. Code

Incorporating the feedback points mentioned above, I will update our AI system code snippet further to include:

1. **Data Governance:** Add methods for managing applicant data with compliance.
2. **Bias Mitigation:** Enhance the bias mitigation strategy.
3. **Explanations:** Implement a simple explanation mechanism.

Here's an updated version of the previous structure:

```
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import cv2  # Placeholder for video processing library

class JobApplicantAI:
    def __init__(self):
        self.resumes = []
        self.video_features = []
        self.vectorizer = TfidfVectorizer()
        self.applicant_data = {}

    def add_applicant(self, applicant_id, resume: str, video_path: str):
        self.resumes.append(resume)
        self.video_features.append(self.extract_video_features(video_path))
        self.applicant_data[applicant_id] = {"resume": resume, "video_path": vic

    def extract_video_features(self, video_path: str):
        # Implement video feature extraction logic using appropriate ML methods
        return np.random.rand(10)  # Simulated feature vector

    def process_resumes(self):
        return self.vectorizer.fit_transform(self.resumes)

    def mitigate_bias(self, scores):
        # Implement bias mitigation strategies, e.g., score adjustment or re-wei
        return scores  # Modify scores to reduce bias (dummy implementation)

    def explain_decision(self, rankings):
        explanations = []
        for idx in rankings:
            explanations.append({
                "applicant_id": idx,
                "justification": "Ranked based on resume matching and video scor
            })
```

Figure 11. *Example of Enhanced Code generated by AI Ethicist for a High-Risk Level AI module description.*

```
        return explanations

    def rank_applicants(self, job_description: str):
        resume_vectors = self.process_resumes()
        job_vector = self.vectorizer.transform([job_description])
        similarity_scores = cosine_similarity(job_vector, resume_vectors)

        combined_scores = similarity_scores.flatten() + np.array(self.video_feat
        adjusted_scores = self.mitigate_bias(combined_scores)

        ranked_indices = np.argsort(adjusted_scores)[::-1]  # Sort descending
        explanations = self.explain_decision(ranked_indices)
        return ranked_indices, explanations

# Example Usage
applicant_ai = JobApplicantAI()
applicant_ai.add_applicant(1, "Resume text of applicant 1", "video1.mp4")
applicant_ai.add_applicant(2, "Resume text of applicant 2", "video2.mp4")
rankings, explanations = applicant_ai.rank_applicants("Desired skills and qualit
print("Rankings:", rankings)
print("Explanations:", explanations)
```

4. Critique

- **Strength:** This code enhancement introduces an explanation mechanism that increases transparency, which is a vital aspect of ethical AI. The addition of a data governance structure indicates our focus on compliance.
- **Areas for Improvement:**
  - **Bias Detection:** We could implement specific techniques to evaluate bias using fairness metrics in the `mitigate_bias` method.
  - **Video Analysis:** Implementing actual feature extraction for video processing, potentially utilizing libraries like OpenCV or deep learning models, remains critical for accurate scoring.

Figure 12. *Example of Enhanced Code (continued) generated by AI Ethicist with Critique.*

EVALUATION AND RESULTS

4.1 Evaluation Setup

To evaluate the ethical utility and usability of the proposed Agents4EthicalSE framework, a workshop-based experiment was conducted with 82 Master's-level students. The evaluation followed a structured survey that combined quantitative Likert-scale items and qualitative open-text responses grounded in practical tasks. Participants were instructed to simulate real-world development workflows as shown in Fig [13], involving:

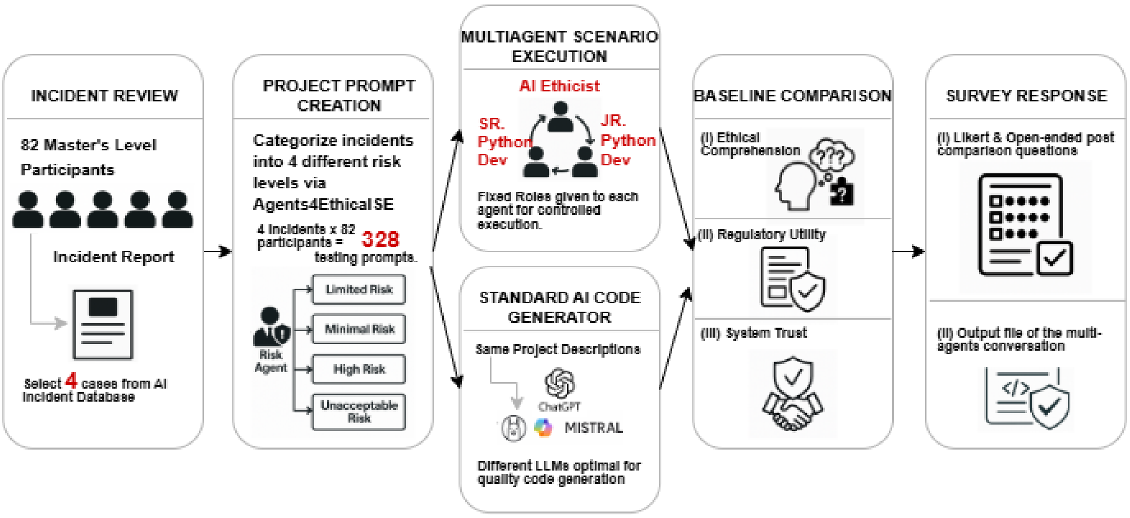

Figure 13. Evaluation Process of Agents4EthicalSE Tool.

Incident review: Each participant selected **four AI incidents** from the [AI Incident Database](), such as algorithmic hiring bias or facial recognition misuse.

Project prompt construction: From each incident, they crafted a project description, categorizing it into one of the following risk levels according to the EU AI Act framework:

1. ***Unacceptable Risk*** – AI systems that threaten fundamental rights.
6. ***High Risk*** – Systems affecting safety or legal entitlements.
7. ***Limited Risk*** – Systems requiring transparency to mitigate potential harm.
8. ***Minimal Risk*** – Systems considered low or no threat.

In total, this process resulted in **328 project descriptions** (4 per participant × 82 participants), each evaluated and classified using the Agents4EthicalSE tool. The system's RiskGuard and AI Ethicist agents analyzed the ethical dimensions and regulatory context of each description before assigning a risk category. This automated classification process supports early-stage ethical awareness and alignment with EU regulations.

Multi-agent scenario execution: Using the Agents4EthicalSE tool, participants assigned agent roles (e.g., *Senior Python Developer*, *Junior Python Developer*) and chose **multi-round or single-round interaction structures**. These agents generated and iteratively critiqued code while applying ethical reasoning.

Baseline comparison: The same prompts were submitted to a **standard AI code generator** (compromising different LLMs i.e., Llama, Mistral, Claude etc.) to compare the outputs' ethical alignment and usability.

Survey completion: Participants then completed a detailed evaluation survey, consisting of **Likert-scale and open-ended items** and attaching the generated output capturing perceptions of ethical guidance, trust, and practical relevance.

4.2 Participant Overview

A total of **82 participants** responded, representing diverse technical backgrounds including **AI research, software engineering, and data science**. Participants reported between **<1 year and 6+ years** of experience in AI-related or software development domains. Most were affiliated with graduate-level programs in **Computer Science** or **Data Science**, aligning well with the expected user base for AI-assisted software tools.

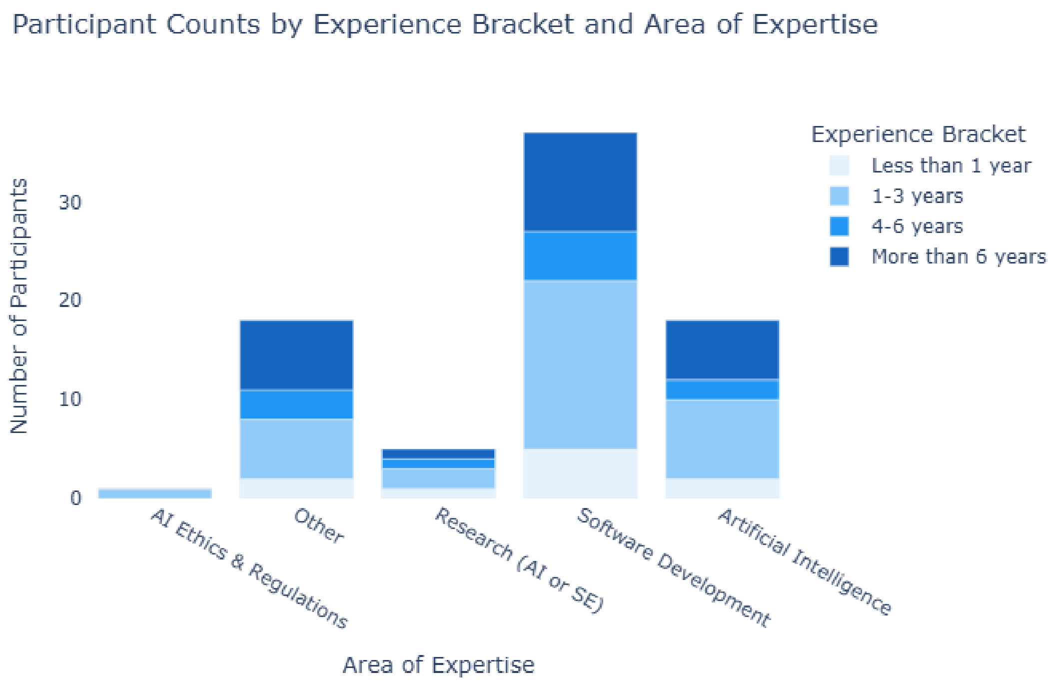

Figure 14. *Participant counts per experience bracket and area of expertise.*

As illustrated in Fig [14]:

1. **Software Development** was the most common discipline (38 participants), with the majority having **1–3 years** (21 participants) or **less than 1 year** (11 participants) of experience.
2. **Artificial Intelligence** had 19 participants, evenly split between early-career (1–3 years and <1 year), indicating a growing AI-exposed student cohort.
3. The **Other** category also had strong representation (19 participants), early-career, suggesting a diverse range of AI-adjacent domains (e.g., HCI, policy, data science).
4. **Research-focused participants** (e.g., those in AI or SE) made up a small share (5 participants), but displayed a wider range of experience, including mid- and late-career individuals.
5. Only one participant reported their primary expertise as **AI Ethics & Regulations**, highlighting the gap between ethical governance experts and those building AI systems.

This demographic distribution reflects a representative group of early-to-mid-career developers' population likely to benefit from assistive ethical tooling in their learning or professional environments.

4.3 Survey Design and Metrics

The survey was designed to operationalize the key evaluation goals using cognitively pretested questions (Martin, 2006) and refined through experts' feedback. Each Likert item was paired with **optional open comment fields** for the inclusion of qualitative insights. It assessed the following dimensions as shown in Table [4]:

Table 4. *Assessment Constructs for Agents4EthicalSE Tool.*

Construct	Purpose	Scale
Ethical Understanding	How well did participants understand concepts like fairness, bias mitigation, explainability?	5-point Likert
Regulatory Utility	Was the tool perceived as helpful for EU AI Act compliance?	5-point Likert
System Trust	Do users trust the system's legal and ethical guidance?	5-point Likert
Recommendation	Would they recommend the tool to others in the field?	5-point Likert

4.4 Quantitative and Qualitative Results

4.4.1 Quantitative Summary

Across the four measured constructs—**ethical comprehension**, **compliance utility**, **system trust**, and **recommendation likelihood**—most participants reported a **positive or moderately positive experience** with Agents4EthicalSE as shown in Fig [15]. The chart clearly shows the varying response patterns across different dimensions:

- **Ethical Comprehension**: Strong positive responses with most participants rating "Significantly" or "Moderately".
- **Regulatory Utility**: Similar pattern to ethical comprehension but with a notable portion (12.2%) rating "Very significantly".
- **Trust in System**: Different pattern with no "Significantly" or "Very significantly" responses, showing more moderate trust levels.
- **Recommendation**: Strong positive responses with a balanced distribution across "Moderately", "Significantly", and "Very significantly".

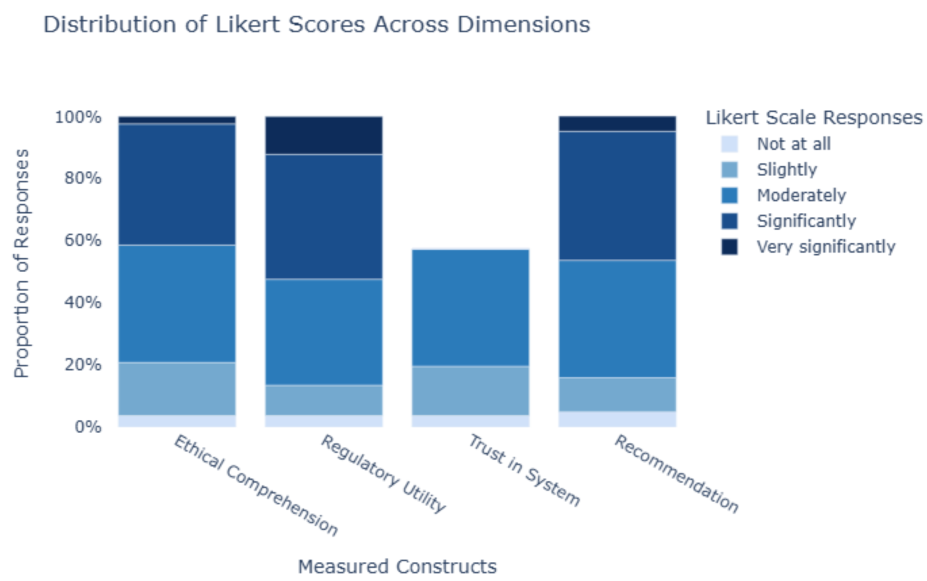

Figure 15. *Distribution of Likert Scores across core Constructs.*

4.4.2 Qualitative Insights

Participants provided free-text responses to describe specific impressions, pain points, and suggestions. Using an inductive coding process, qualitative feedback was analyzed to extract recurring themes and patterns from participants' comments as highlighted in Table [5].

Table 5. *Thematic Analysis of Participants' Feedback.*

Theme Category	Content	Reprsentative Quote	Frequency
Positive Insight	Comments about the tool's ethical capabilities and benefits.	*- "The AI ethicist agent could provide references to regulatory guidelines to enhance user trust."* *- "I think it is a good initiative to bring AI to SE. It will help a lot of developers as well as the companies."*	High
Instructional Value	Feedback on the learning and reflection process.	*- "Multiple rounds helped me reflect more critically."* *- "Using this system, I believe that mitigation strategies and risk anticipating could be applied in real-world setting."*	Moderate
Citation Concerns	Issues related to transparency and traceability of citations.	*- "I think the system could be more transparent and provide the direct source where in the legislation it is getting its categorization of a project."* *- "I found the tool pretty handy. Maybe the references given by the AI ethicist could be even more specific."*	High
Coordination Issues	Observations about agent role clarity and response overlap.	*- "Some agent responses overlapped—could use more clear role separation."* *- "Right now, I feel like it did not use any information generated before. Junior Developer didn't use information of Senior Developer."*	Moderate
Suggestions	Recommendations for tool improvement and additional features.	*- "Expanded Agent Ecosystem: Include specialized roles like Security Expert or UX Designer."* *- "I would integrate the system into existing developer tools such as IDEs."*	Moderate

The theme frequency analysis (Fig [16]) reveals:

- **Positive Insight (High Frequency)**: Participants noted that the AI Ethicist could provide references to regulatory guidelines, which would enhance trust.
- **Instructional Value (Moderate Frequency)**: Multiple rounds of interaction helped participants reflect more critically on ethical issues.
- **Citation Concerns (High Frequency)**: A common concern was the lack of clear citation transparency and traceability.
- **Coordination Issues (Moderate Frequency)**: Some participants observed overlapping agent responses, suggesting a need for clearer role separation.
- **Suggestions (Moderate Frequency)**: Participants recommend adding more specialized agent roles, such as policy advisors or QA engineers.

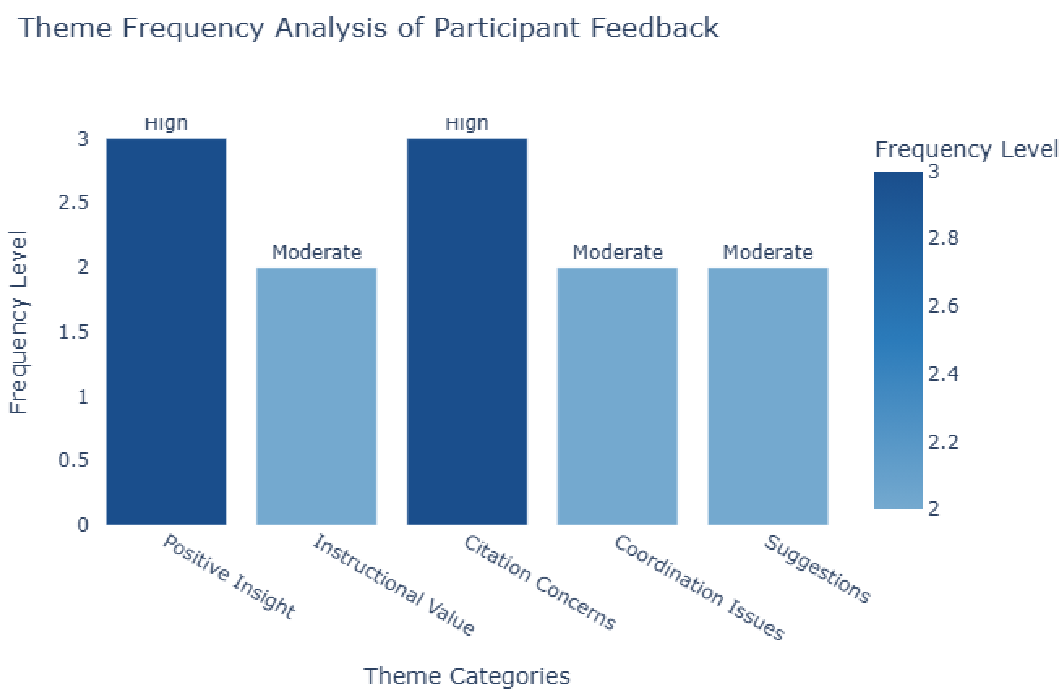

Figure 16. *Theme Frequency Analysis of Participants' Feedback.*

4.5 Key Findings per Research Question

RQ1: How do agent specialization, multi-round interactions, structured prompts, and conversational protocols contribute to the ethical alignment of outputs?

Participants consistently reported that the **structured agent roles** (e.g., *AI Ethicist* vs *Senior Python Developer*) helped make ethical issues **more explicit and actionable**. The multi-round structure encouraged **critical reassessment**, enabling users to revise assumptions—like the **reflection cycles** emphasized in ethical design thinking.

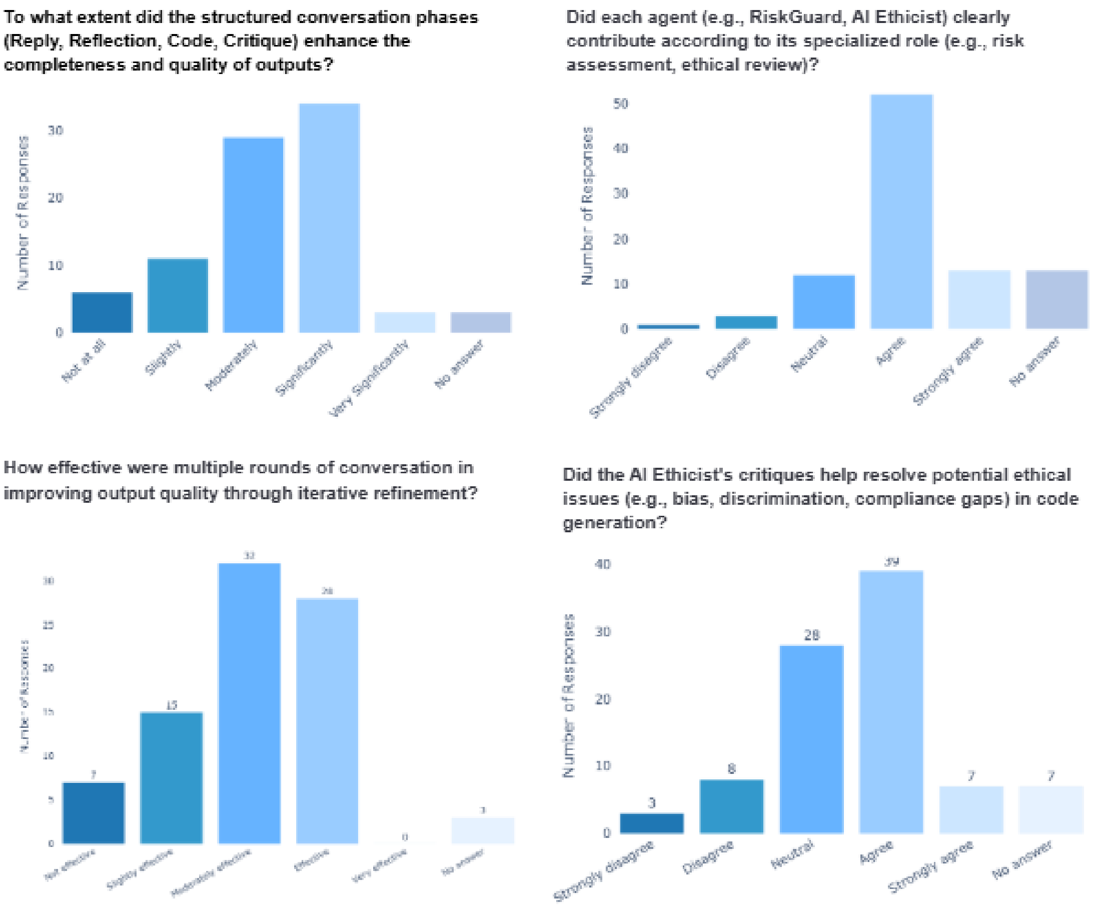

Figure 17. *Participants' perception of agent specialization, multi-round interactions, structured prompts, and conversational protocols.*

The findings in Fig [17] affirm the design logic that **role-specific critique combined with conversational sequencing** promotes ethical deliberation, aligning with the EU AI Act's emphasis on accountability and transparency.

RQ2: In what ways can a Multi-Agent RAG framework assist in aligning LLMs with established ethical standards and regulatory guidelines?

The system's use of **retrieval-augmented generation (RAG)** contributed to participant trust (*see Fig [18]*), especially when the agents referenced actual legal texts (e.g., HLEG guidelines or EU AI Act provisions). However, a recurring concern was **lack of traceability** in 15% citations.

Participants valued:

- Context-aware legal guidance
- Role-driven ethical critique
- But also requested: "Show paragraph numbers or links in citations."

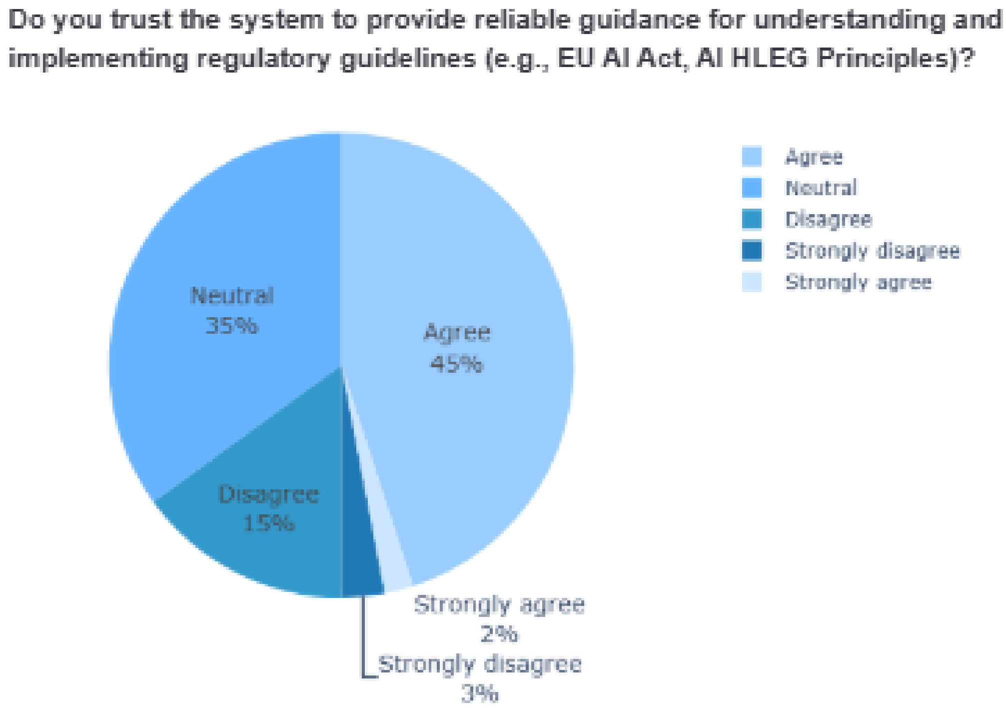

Figure 18. *Participants' perception of reliable regulatory guidance via RAG*

This supports the broader literature suggesting that **explainability + traceability** form the foundation of trustworthy AI.

4.6 Reflections and Limitations

This evaluation validates the **promise of multi-agent RAG ethical reasoning during software development**. The integration of structured prompts, agent specialization, and real-world AI incident simulation made the experience authentic and pedagogically rich.

That said, limitations include:

- **Construct validity**: Self-reported learning may not equate to improved ethical outcomes in practice.
- **External validity**: Academic participants may not reflect high-pressure industry settings.
- **System friction**: Interface lags and vague RAG references occasionally undermined user experience.

Despite these limitations, the findings suggest Agents4EthicalSE holds promise for **augmenting ethical compliance efforts** in both educational and professional settings.

THREAT TO VALIDITY

This section discusses potential threats that may influence the reliability, generalizability, or interpretability of the results presented in this thesis.

5.1 Internal Validity

Construct validity refers to whether the concepts studied (e.g., “ethical alignment” or “agent deliberation”) are appropriately operationalized. While the study defines alignment as outputs consistent with ethical principles and regulations, there is inherent ambiguity in how various sources define ethics (e.g., EU AI Act vs. OECD principles). Although retrieval-augmented grounding reduces vagueness, the framework's judgment is still dependent on prompt structure and source quality.

5.2 Construct Validity

The internal consistency of the Agents4EthicalSE evaluation may be affected by agent behavior variance. LLMs can produce inconsistent responses depending on temperature settings, prompt phrasing, or length. While the system uses deterministic settings (temperature = 0) and few-shot examples for calibration, unexpected model deviations remain possible. In addition, the multi-round interaction protocol was designed to simulate deliberation but may not guarantee epistemic convergence.

5.3 External Validity

External validity is limited due to the academic setting of the evaluation. Although real-world ethics guidelines and regulatory texts were used, the decision scenarios and system walkthroughs were hypothetical or educational. Further deployment in industry settings (e.g., agile retrospectives, code audits) is needed to validate generalizability. The framework also assumes a baseline understanding of ethical concerns, which may vary across software teams and regions.

5.4 Reliability

Agent outputs are reproducible given the same input conditions, but performance can vary across models or API versions. No systematic stress-testing or adversarial prompting was performed, which may affect robustness claims.

DISCUSSION

This chapter interprets the findings of the thesis through five thematic lenses. It connects the design and evaluation of the Agents4EthicalSE framework to broader challenges in AI ethics, LLM deployment, and software engineering support tools.

6.1 Ethical Alignment through Agent Roles and Dialogue

The structured multi-agent architecture implemented in Agents4EthicalSE demonstrates how ethical deliberation can be constructed through agent specialization and multi-round feedback. Drawing from the literature review (Section 3.2), agent role assignment—such as “RiskGuard” for value critique or “AIEthicist” for regulatory interpretation—mirrors real-world role separation in ethics committees or software quality audits.

Studies like Zeng et al. [14] and Xu et al. [37] emphasize that distributing responsibility across role-specialized agents can increase reasoning quality, especially when dialogue is iterated. In the implemented framework, multi-turn interactions proved valuable for surfacing trade-offs (e.g., between accuracy and privacy), enabling a richer understanding of complex ethical tensions. This suggests that internal deliberation, even if artificial, can guide alignment when ethical ambiguity arises.

6.2 Regulation-Aware Development via RAG

The integration of Retrieval-Augmented Generation (RAG) enables context-sensitive decision-making by grounding agent discussions in specific regulatory documents such as the EU AI Act, ISO/IEC TR 24028, and organizational ethics policies. This reflects a key insight from both Lewis et al. [15] and more recent compliance-focused works such as Kővári et al. (2025): AI agents are more trustworthy when they cite interpretable sources.

Unlike static rule-based checklists, Agents4EthicalSE retrieves relevant passages at runtime, offering just-in-time legal awareness. This not only reduces hallucinations but also enables real-time traceability—a critical concern for explainability and auditability in high-risk systems. RAG thereby acts as an “external conscience” for the agent ensemble.

6.3 Developer Perceptions and Practical Value

Preliminary evaluations through controlled workshop and developer walkthroughs suggest the tool offers perceived value across three dimensions: (1) enhanced ethical awareness, (2) practical assistance in documentation, and (3) improved engagement with regulatory requirements. Participants reported that the AIEthicist's retrieval-backed critiques improved their understanding of legal risks.

However, developers also highlighted that the conversational format—while educational—could be overly verbose for fast-paced development environments. This points to a trade-off between depth and usability, and suggests that lightweight modes (e.g., summaries or flagging-only agents) may be desirable for production settings.

6.4 Comparison with Existing Tools

Compared to existing tools such as ethics canvases, risk taxonomies, or fairness dashboards, Agents4EthicalSE introduces two innovations: multi-agent dialogue and dynamic legal grounding. Tools like IBM's AI FactSheets or Google's Model Cards offer post-hoc documentation but lack deliberative reasoning. Conversely, RiskGuard and AIEthicist agents simulate argumentation during the design phase, promoting anticipatory reflection.

Moreover, while some frameworks use static rule encoding (e.g., rulebooks or ethical matrices), this system adapts its reasoning based on retrieved context. In this sense, it aligns with "ethics-as-a-service" paradigms [8] but advances them through a modular, explainable, and interactive interface.

6.5 Barriers to Adoption

Despite its promise, several barriers may limit the adoption of such a framework in industry:

Scalability: Multi-round conversations may be too time-consuming for agile workflows unless optimized or selectively triggered.

Domain Adaptation: The framework currently depends on general regulatory texts. Domain-specific fine-tuning (e.g., healthcare, education) is needed for accurate critique.

Trust and Transparency: Developers may distrust opaque decisions by AI agents. Although RAG enhances explainability, the reasoning steps of LLMs remain partly inscrutable.

Integration Complexity: Embedding the tool into real-world pipelines (e.g., GitHub Actions, Jira workflows) requires engineering effort and alignment with DevOps practices.

Addressing these barriers will require iterative co-design with practitioners, user-centric evaluation, and simplification of interaction models for low-friction use.

CONCLUSIONS

This chapter summarizes the key findings of the study, answers the research questions, and reflects on the broader implications of the proposed Agents4EthicalSE framework. It also outlines potential directions for future work.

7.1 Summary of Findings

This thesis investigated how multi-agent RAG systems can support ethical software development through both a systematic literature review and the design of a retrieval-augmented deliberation tool. The findings reveal:

- There is a growing body of peer-reviewed research on agent collaboration, structured prompting, and dialogue governance in ethical AI, but limited practical implementations exist for software engineering contexts.
- The proposed multi-agent framework—*Agents4EthicalSE*—integrates role-specialized agents, multi-round deliberation, and RAG-based grounding to simulate ethical review and regulatory assessment.
- Initial evaluations suggest that the framework can enhance ethical awareness, support regulation-aware reasoning, and improve transparency in design decisions.

7.2 Answers to Research Questions

RQ1: *How do agent specialization, multi-round interactions, structured prompts, and conversational protocols contribute to the ethical alignment of outputs generated by a multi-agent LLM system?*

The systematic literature review demonstrated that these four features are critical for improving ethical alignment in LLM outputs. Agent specialization introduces diverse perspectives; multi-round dialogue enables deeper reasoning and self-correction; structured prompts constrain outputs toward desirable norms; and conversational protocols enhance transparency and deliberative structure.

RQ2: *In what ways can a Multi-Agent RAG framework assist in aligning LLMs with established ethical standards and regulatory guidelines?*

The developed tool shows that RAG-enabled agents can retrieve and apply up-to-date regulatory knowledge (e.g., EU AI Act), and structure deliberation around it. The integration of legal grounding, agent-based debate, and traceable interactions offers a promising co-pilot model for ethics-by-design in AI development workflows.

7.3 Contributions and Implications

This work contributes to the fields of AI ethics, software engineering, and human-AI interaction in distinct ways:

- **Empirical synthesis**: It presents the systematic mapping of peer-reviewed research linking four design strategies (agent roles, iterations, prompts, dialogue) to ethical alignment in LLMs.
- **Framework design**: It introduces the *Agents4EthicalSE* architecture, combining multi-agent deliberation with retrieval-augmented generation, implemented using modern LLM orchestration tools (OpenAI API).
- **Practice relevance**: It proposes a replicable structure for aligning software artifacts with ethical and legal expectations, offering value for compliance teams, educators, and software architects.

The implications are broad: the system could be adapted for high-risk AI scenarios, embedded into design sprints, or extended to policy compliance support. It positions AI as both subject to ethical critique and a participant in ethical discourse.

7.4 Future Work and Recommendations

Promising directions for future work emerged from both theoretical analysis and practitioner feedback:

- **Human-in-the-loop refinement**: Integrating manual review checkpoints into the agent dialogue process could increase reliability and trustworthiness. For instance, agents might pause to request developer approval before finalizing recommendations.
- **Domain adaptation**: The framework currently uses general-purpose regulatory corpora. Future work should explore domain-specific fine-tuning (e.g., for medical, legal, or educational applications) to tailor ethical scrutiny to local needs.
- **Scalability and integration**: A particularly valuable suggestion emerged from participant feedback. One respondent highlighted the importance of building the Agents4EthicalSE framework as an **in-built plugin for Integrated Development**

Environments (IDEs). This idea reflects a fundamental direction for future work: to embed ethical deliberation into the developer's workflow without requiring a separate interface or toolchain. Integration with environments like Visual Studio Code, IntelliJ, or GitHub Copilot could allow agents to proactively evaluate code decisions, offer ethical suggestions inline, or flag regulatory risks during commit.

- **Usability optimization**: The verbosity and multi-turn nature of the current dialogue format may not suit all workflows. Future designs could explore compressed dialogue formats, agent role toggling, and adaptive levels of explanation depending on user preference.
- **Cross-cultural and multilingual generalization**: To ensure fairness and inclusivity, future extensions should incorporate region-specific policies, language diversity, and culturally informed ethical interpretations.

By pursuing these directions—particularly the IDE plugin pathway—Agents4EthicalSE can evolve from a research prototype into a usable co-pilot for ethically aware software engineering.

REFERENCES

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019. doi: 10.1038/s42256-019-0088-2.

[2] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, vol. 1, no. 1, Jul. 2019. doi: 10.1162/99608f92.8cd550d1.

[3] L. Floridi, "Soft Ethics and the Governance of the Digital," *Philosophy & Technology*, vol. 31, pp. 1–8, 2018, doi: 10.1007/s13347-018-0303-9.

[4] J. Morley, L. Kinsey, A. Elhalal, and L. Floridi, "Operationalising AI ethics: barriers, enablers and next steps," *AI & Society*, vol. 38, pp. 411–423, 2023, doi: 10.1007/s00146-021-01308-8.

[5] R. Vidgen, G. Hindle, and I. Randolph, "Exploring the ethical implications of business analytics with a business ethics canvas," *European Journal of Operational Research*, vol. 281, no. 3, pp. 491–501, Dec. 2019, doi: 10.1016/j.ejor.2019.03.004.

[6] V. Vakkuri, K.-K. Kemell, J. Kultanen, and P. Abrahamsson, "The current state of industrial practice in artificial intelligence ethics," *IEEE Software*, vol. 37, no. 4, pp. 50–57, Jul.–Aug. 2020, doi: 10.1109/MS.2020.2985621.

[7] M. Agbese, R. Mohanani, A. A. Khan, and P. Abrahamsson, "Ethical Requirements Stack: A framework for implementing ethical requirements of AI in software engineering practices," in *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering (EASE '23)*, Oulu, Finland, Jun. 2023, pp. 326–328, doi: 10.1145/3593434.3593489.

[8] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mokander, and L. Floridi, "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics," *Minds and Machines*, vol. 31, pp. 239–256, 2021, doi: 10.1007/s11023-021-09563-w.

[9] Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, and X. Qiu, "Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-Model Communication," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Singapore: Assoc. for Computational Linguistics, 2023, pp. 15135–15153. doi: 10.18653/v1/2023.emnlp-main.936.

[10] J. C.-Y. Chen, S. Saha, and M. Bansal, "ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs," *arXiv preprint*, arXiv:2309.13007, 2024. [Online]. Available: http://arxiv.org/abs/2309.13007

[11] J. C. Yang, D. Dailisan, M. Korecki, C. I. Hausladen, and D. Helbing, “LLM Voting: Human Choices and AI Collective Decision Making,” *arXiv preprint*, arXiv:2402.01766, 2024. [Online]. Available: http://arxiv.org/abs/2402.01766

[12] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving Factuality and Reasoning in Language Models through Multiagent Debate,” *arXiv preprint*, arXiv:2305.14325, 2023. [Online]. Available: http://arxiv.org/abs/2305.14325

[13] Q. Sun, Z. Yin, X. Li, Z. Wu, X. Qiu, and L. Kong, “Corex: Pushing the Boundaries of Complex Reasoning through Multi-Model Collaboration,” *arXiv preprint*, arXiv:2310.00280, 2024. [Online]. Available: http://arxiv.org/abs/2310.00280

[14] Y. Zeng, Y. Wu, X. Zhang, H. Wang, and Q. Wu, “AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks,” *arXiv preprint*, arXiv:2403.04783, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.04783

[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proc. 34th Int. Conf. Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 9459–9474.

[16] V. Vakkuri, K.-K. Kemell, and P. Abrahamsson, “ECCOLA – a Method for Implementing Ethically Aligned AI Systems,” in *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Portoroz, Slovenia, Aug. 2020, pp. 195–204. doi: 10.1109/SEAA51224.2020.00043.

[17] J. S. de Cerqueira, A. Azevedo, H. Tives, and E. D. Canedo, “Guide for Artificial Intelligence Ethical Requirements Elicitation – RE4AI Ethical Guide,” in *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2022. doi: 10.24251/HICSS.2022.677.

[18] C. Sanderson, D. Douglas, Q. Lu, E. Schleiger, J. Whittle, J. Lacey, G. Newnham, S. Hajkowicz, C. Robinson, and D. Hansen, “AI Ethics Principles in Practice: Perspectives of Designers and Developers,” CSIRO, Australia, Tech. Rep., 2020.

[19] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices,” *Science and Engineering Ethics*, vol. 26, 2020. doi: 10.1007/s11948-019-00165-5.

[20] Y. Liu, J. Cao, Z. Li, R. He, and T. Tan, “Breaking Mental Set to Improve Reasoning through Diverse Multi-Agent Debate,” in *Proc. 13th Int. Conf. on Learning Representations (ICLR)*, 2025. [Online]. Available: https://openreview.net/forum?id=t6QHYUOQL7

[21] L. Reynolds and K. McDonell, “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,” *arXiv preprint* arXiv:2102.07350, 2021. [Online]. Available: https://arxiv.org/abs/2102.07350

[22] Y. Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” *arXiv preprint* arXiv:2212.08073, 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[23] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” EBSE Technical Report, 2007.

[24] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004. doi: 10.2307/25148625.

[25] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “Design Science Research Methodology Process Model,” in *Design Science Research Methodology Process Model according to Peffers et al.*, C. Krupitzer, Ed. [Online]. Available: https://www.researchgate.net/profile/Christian-Krupitzer/publication/340262762/figure/fig3/AS:874395346157570@1585483504410

[26] G. Thomas and D. Harden, “Methods for the thematic synthesis of qualitative research in systematic reviews,” BMC Medical Research Methodology, vol. 8, no. 1, p. 45, 2008. doi: 10.1186/1471-2288-8-45.

[27] C. Yin, F. Li, S. Zhang, Z. Wang, J. Shao, P. Li, et al., “MDD-5k: A New Diagnostic Conversation Dataset for Mental Disorders Synthesized via Neuro-Symbolic LLM Agents,” Proceedings of the AAAI Conference on Artificial Intelligence, 2025. DOI: 10.1609/aaai.v39i24.34763.

[28] F.R.E. Silva, P.A. Santos, J. Dias, “MentalRAG: Developing an Agentic Framework for Therapeutic Support Systems,” Proceedings of the International Conference on ICT for Ageing Well and e-Health (ICT4AWE), 2025. DOI: 10.5220/0013267400003938.

[29] A. Bandi, H. Noh, N. Gomathi, D.S. Mamillapalli, H.P. Gurram, B.P. Bathini, *“Conversational AI in Healthcare: A Framework for Privacy, Security, Ethics, Transparency and Harm Prevention,”* **Communications in Computer and Information Science (Springer)**, 2025. DOI: 10.1007/978-3-031-92178-0_4.

[30] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, et al., “JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models,” Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024. https://arxiv.org/abs/2404.01318

[31] J. Yu, J. Zhao, L. Miranda-Moreno, M. Korp, “Modular AI Agents for Transportation Surveys and Interviews: Advancing Engagement, Transparency, and Cost Efficiency,” Communications in Transportation Research, 2025. DOI: 10.1016/j.commtr.2025.100172.

[32] C.J. Lynch, E.J. Jensen, V. Zamponi, K. O’Brien, E. Frydenlund, R. Gore, “A Structured Narrative Prompt for Prompting Narratives from Large Language Models: Sentiment Assessment of ChatGPT-Generated Narratives and Real Tweets,” Future Internet, 2023. DOI: 10.3390/fi15120375.

[33] S. Lim, R. Schmälzle, “Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering,” Frontiers in Communication, 2023. DOI: 10.3389/fcomm.2023.1129082.

[34] C. Djeffal, "Reflexive Prompt Engineering: A Framework for Responsible Prompt Engineering and AI Interaction Design," Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2025. DOI: 10.1145/3715275.3732118.

[35] M. Ariely, A. Salman, A. Yarden, G. Alexandron, "Reflective Prompt Engineering: A New Strategy for Automated Short Answer Scoring in Biology," International Journal of Science Education, 2025. DOI: 10.1080/09500693.2025.2523571.

[36] T. Wang, Y. Zhan, J. Lian, Z. Hu, N.J. Yuan, Q. Zhang, et al., "LLM-Powered Multi-Agent Framework for Goal-Oriented Learning in Intelligent Tutoring System," Companion Proceedings of the ACM Web Conference (WWW 2025), 2025. DOI: 10.1145/3701716.3715244.

[37] Z. Xu, W. Chen, Y. Tang, X. Li, C. Hu, Z. Chu, et al., "Mitigating Social Bias in Large Language Models: A Multi-Objective Approach Within a Multi-Agent Framework," Proceedings of the AAAI Conference on Artificial Intelligence, 2025. DOI: 10.1609/aaai.v39i24.34748.

[38] M. Vaccaro Jr., M. Friday, A. Zaghi, "Multi-Agentic LLMs for Personalizing STEM Texts," Applied Sciences (Switzerland), 2025. DOI: 10.3390/app15137579.

[39] K. Sakaguchi, R. Sakama, T. Watari, "Evaluating ChatGPT in Qualitative Thematic Analysis with Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study," Journal of Medical Internet Research, 2025. DOI: 10.2196/71521.

[40] H. Liu, Y. Lei, Z. Wu, *"Prosocial Behavior in Large Language Models: Value Alignment and Affective Mechanisms,"* **Science China Technological Sciences**, 2025. DOI: 10.1007/s11431-025-2977-0.

[41] A.M. Goloujeh, A. Sullivan, B. Magerko, *"The Social Construction of Generative AI Prompts,"* **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**, 2024. DOI: 10.1145/3613905.3650947.

[42] M. Bearman, R. Ajjawi, *"Learning to Work with the Black Box: Pedagogy for a World with Artificial Intelligence,"* **British Journal of Educational Technology**, 2023. DOI: 10.1111/bjet.13337.

[43] F.S. Malik, O. Terzidis, *"A Hybrid Framework for Creating Artificial Intelligence-Augmented Systematic Literature Reviews,"* **Management Review Quarterly**, 2025. DOI: 10.1007/s11301-025-00522-8.

APPENDIX A:

1. Systematic Literature Review Log Sheet.xlsx: This sheet contains: Titles, authors, and publication metadata of all retrieved studies. Inclusion and exclusion decisions with justifications. Screening outcomes for each stage (title, abstract, full text). Notes from reviewers.
2. Papers Contributing to 4 major themes.xlsx: This sheet includes: Detailed mapping of each study to one or more of the four core design features: Agent specialization, multi-round interactions, structured prompting, conversational protocols.
3. Agents4EthicalSE Conversation History: An example conversation output file that was generated by a participant during the system evaluation phase of the study. The file includes full conversation logs between agents and output messages in raw text format.
4. Survey.csv: This sheet contains anonymous participants' responses to the workshop survey conducted to evaluate the effectiveness of Agents4EthicalSE tool.

APPENDIX B : SOURCE CODE REPOSITORY

This appendix provides access to the **private GitHub repository** *(accessible to GPT-Lab members due to publication reasons)* that contains the source code, scripts, and documentation associated with the development of the **Agents4EthicalSE** system described in this thesis.

The repository includes:

- Source code for the tool includes frontend and backend, vector store creation, CRUD actions for agents, agents' coordination, and task execution.
- Prompt templates and configuration files.
- README documentation for setup and usage.

🔗 **GitHub Repository**:

https://github.com/GPT-Laboratory/Agents4EthicalSE

Note: *This repository is privately accessible and hosted on GitHub. Any modifications or usage should follow the license terms provided in the repository.*