https://doi.org/10.64504/big.d.v3i3.790

Leveraging Artificial Intelligence for Ethical Design Decision-Making: Philosophical Perspectives on Technical Integration, Value Proposition, and Implementation Challenges

1st Lingyan Zhang*
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
zhlingyan@zju.edu.cn

2nd Xusheng Zhang
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
zhangxs001@zju.edu.cn

Received on April 7th; revised on May 10th, accepted on May 20th, published on July 6th

***Abstract*—Ethical considerations are paramount in modern engineering design, yet designers often lack systematic tools for navigating complex ethical dilemmas. This paper proposes an AI-driven system, the Ethical Deliberation Assistant (EDA), to enhance ethical decision-making in design. The EDA integrates advanced technical strategies, including prompt engineering, Retrieval-Augmented Generation (RAG), and fine-tuning of Large Language Models (LLMs), with established ethical frameworks. Our technical framework details the system architecture and implementation, focusing on quantifiable metrics for ethical assessment. Experimental evaluation demonstrates that the EDA significantly improves the consistency and accuracy of ethical evaluations, achieving an 86% accuracy rate compared to expert judgments and reducing analysis time by 90%. This system provides structured ethical guidance, broadens access to ethical expertise, and fosters a more reflective and efficient design practice, addressing critical engineering challenges in responsible AI development.**

***Keywords*—Design Ethics, Artificial Intelligence, Ethical Decision-Making, Explainable AI (XAI), Retrieval-Augmented Generation, Engineering**

1. Introduction

The pervasive integration of technology into daily life has amplified the ethical implications of engineering design. Designers are increasingly confronted with complex challenges, often termed "wicked problems," that necessitate rigorous ethical deliberation, encompassing issues such as user privacy, algorithmic fairness, data security, and environmental sustainability [1]. Traditional ethical decision-making approaches, relying on professional codes or ad-hoc discussions, are often insufficient to address the novel and rapidly evolving ethical landscape presented by emerging technologies, particularly Artificial Intelligence (AI). The potential for widespread negative consequences from even minor design flaws in digital products underscores the urgent need for systematic and rigorous ethical frameworks within the design process.

The rapid advancements in Large Language Models (LLMs) offer unprecedented opportunities to augment human cognitive capabilities, including complex ethical reasoning. This paper investigates the technical integration of LLMs into the ethical design process, proposing an AI-driven Ethical Deliberation Assistant (EDA). The EDA is designed as a sophisticated decision-support tool, leveraging prompt engineering, Retrieval-Augmented Generation (RAG), and fine-tuning techniques to provide context-aware ethical analysis and guidance. This approach aims to bridge the gap between abstract ethical principles and the concrete, practical challenges faced by engineering designers, thereby enhancing the quality and consistency of ethical considerations in technological development.

This research addresses the following key engineering and technical questions:

- RQ1: How can LLMs be technically integrated into the design workflow to effectively assist in ethical decision-making, ensuring scalability and maintainability?
- RQ2: What is the quantifiable value proposition of an LLM-assisted system for improving the quality, consistency, and efficiency of ethical design practices?
- RQ3: What are the primary technical challenges and ethical considerations that must be addressed in the development and deployment of such a system, particularly regarding algorithmic bias and transparency?

2. Related Work

The evolution of design ethics has shifted from a focus on professional duties to a broader examination of the societal and environmental consequences of design,

*Lingyan Zhang, College of Computer Science and Technology, Zhejiang University, Hangzhou, China, zhlingyan@zju.edu.cn

particularly with the advent of digital technologies and AI [2]. This section reviews existing literature on design ethics, AI ethics, and decision support systems, emphasizing engineering and technical perspectives to identify current gaps and contextualize our proposed solution.

2.1. Design Ethics and AI Integration

Traditional design ethics, influenced by thinkers like Victor Papanek, advocated for design practices addressing genuine human needs and social justice. Richard Buchanan's concept of "wicked problems" highlighted the inherent ethical complexity in design, where solutions often create new problems. The rise of digital technologies has introduced new ethical challenges, including data privacy, algorithmic bias, and digital well-being, making design ethics an interdisciplinary field drawing from philosophy, sociology, law, and computer science [3]. Recent work has explored frameworks for responsible innovation, but often lacks concrete technical implementation strategies for integrating ethical considerations directly into the design workflow.

Integrating AI into design processes has been a growing area of research. Zhou et al. [4] examined how LLMs support the entire design process, from discovery to delivery, highlighting various functions. However, their work primarily focuses on the functional aspects rather than the ethical decision-making support. Esmaeilzadeh [5] discussed the ethical implications of using general-purpose LLMs in clinical decision-making, raising concerns about patient safety and reliability, which are analogous to design contexts. These studies underscore the need for robust ethical frameworks when deploying AI in critical decision-making scenarios.

2.2. AI Ethics and Quantifiable Assessment

AI ethics has rapidly emerged as a critical field, focusing on ensuring fairness, transparency, accountability, and safety in AI systems. Jiao et al. [1] proposed a three-dimensional assessment system for evaluating moral reasoning in LLMs, quantifying alignment with human ethical standards through foundational moral principles, reasoning robustness, and value consistency. This work provides a crucial benchmark for evaluating the ethical capabilities of LLMs, moving beyond qualitative assessments to measurable metrics. Similarly, Hadar-Shoval et al. [6] investigated how embedded values-like profiles within LLMs impact their ethical reasoning, suggesting that intrinsic model characteristics influence ethical outputs.

Quantitative evaluation of AI ethics is gaining traction. Ferdaus et al. [7] provided a comprehensive review of ethical and robust LLMs, including quantitative evaluation results across various trustworthiness dimensions. Awad et al. [8] discussed computational ethics, emphasizing the use of logical, mathematical formulas, and computational models to quantify moral intuitions. These studies highlight the shift towards a more rigorous, data-driven approach to AI ethics, which is essential for developing trustworthy AI systems. However, a direct application of these quantitative ethical assessment methods to real-world engineering design decision-making processes, particularly with LLM-assisted systems, remains an area requiring further exploration.

2.3. Decision Support Systems and LLM Integration

Decision Support Systems (DSS) have long been utilized to aid human decision-making in complex environments. The integration of LLMs into DSS has shown promising results across various domains. Alkayyal [9] introduced StrategicAI, an LLM-based DSS for strategic decision-making in business environments, demonstrating the potential of LLMs to embed structured frameworks for root cause analysis and solution generation. Arif et al. [10] proposed an AI-Driven Decision Support System (AIDSS) for software architecture, leveraging AI to assist architects in the design stage and using Architecture Decision Records (ADRs) to assess the performance of LLM-based suggestion modules.

Retrieval-Augmented Generation (RAG) has emerged as a powerful technique to enhance LLM performance by integrating external knowledge retrieval, thereby addressing issues of factual inconsistency and knowledge limitations [11]. Pradhan [12] provided a comparative analysis of prompt engineering, fine-tuning, and RAG for optimizing AI models, emphasizing their respective strengths in different use cases. Wang et al. [13] and Wan et al. [14] explored hybrid RAG frameworks for improving knowledge management in building engineering and domain-centric Q&A in smart manufacturing, respectively, showcasing the engineering value of RAG in specialized knowledge domains. A comprehensive survey by Gan et al. [15] detailed RAG evaluation methods and frameworks, systematically reviewing traditional and emerging approaches for system performance, factual accuracy, safety, and computational efficiency. Collaco [16] further demonstrated that FT + RAG systems consistently outperformed FT-only or RAG-only approaches across various decision support tasks. These works collectively establish the technical feasibility and benefits of integrating LLMs and RAG into DSS, providing a strong foundation for our proposed EDA system.

2.4. Research Gap

While significant progress has been made in AI ethics and LLM-based decision support, a critical gap exists in the systematic engineering and quantitative evaluation of LLM-assisted ethical decision-making systems specifically tailored for design processes. Existing literature often focuses on either philosophical aspects of design ethics or general AI ethics benchmarks. There is a lack of comprehensive frameworks that detail the technical integration of LLMs (using techniques like RAG and fine-tuning) into a practical design workflow, coupled with robust, quantifiable metrics to assess their effectiveness, consistency, and engineering value in real-world design scenarios. Our work aims to bridge this gap by presenting a technically detailed and experimentally validated EDA system.

3. Methodology and System Design

This section outlines the technical methodology and system design of the Ethical Deliberation Assistant (EDA), focusing on its architecture, core AI techniques, and the integration of ethical frameworks. The EDA is engineered to provide quantifiable ethical guidance, ensuring reproducibility and systematic application within design workflows.

3.1. System Architecture

The EDA system is designed with a modular, three-layered architecture to ensure scalability, maintainability, and clear separation of concerns, as illustrated in Figure 1.

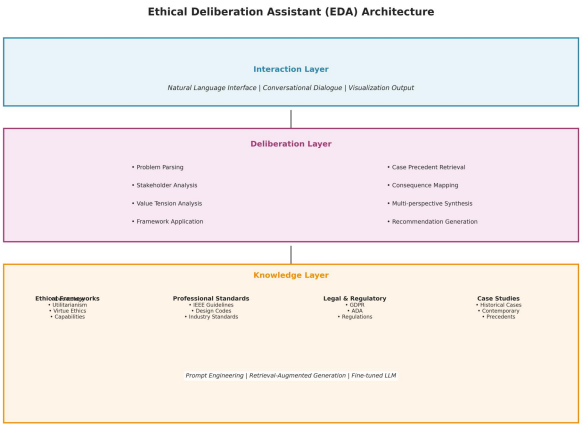

Figure 1. Ethical Deliberation Assistant (EDA) Architecture

These layers are:

- Interaction Layer: This layer provides the user interface for designers to input design problems, receive ethical analyses, and interact with the system. It handles user queries, displays results, and facilitates feedback mechanisms. This layer is responsible for translating designer input into structured queries for the Deliberation Layer and presenting complex ethical insights in an understandable format.
- Deliberation Layer: This is the cognitive core of the EDA, responsible for processing ethical queries, applying ethical frameworks, and generating ethical analyses. It leverages LLMs augmented with specialized knowledge. This layer orchestrates the prompt engineering, RAG, and fine-tuning processes to produce relevant and coherent ethical guidance.
- Knowledge Layer: This layer serves as the grounding for ethical reasoning. It comprises a curated knowledge base of ethical principles, case studies, philosophical frameworks, and design ethics guidelines. This layer is continuously updated and refined to ensure the LLM has access to authoritative and contextually relevant ethical information.

3.2. Core AI Techniques

The EDA system employs a combination of advanced LLM techniques to achieve robust ethical reasoning capabilities:

3.2.1. Prompt Engineering

Prompt engineering is crucial for guiding the LLM to perform specific ethical reasoning tasks. We utilize structured prompts that define the role of the LLM (e.g., ethical advisor), specify the ethical framework to be applied (e.g., utilitarianism, deontology), and outline the desired output format (e.g., identification of ethical dilemmas, stakeholder impact analysis, proposed solutions). For instance, a prompt might include:

"As an ethical design advisor, analyze the following design scenario from a utilitarian perspective. Identify potential ethical dilemmas, quantify stakeholder impacts (positive and negative), and propose design modifications to maximize overall well-being. Scenario: [Design Scenario Description]"

This structured approach ensures that the LLM focuses on relevant ethical dimensions and generates actionable insights, moving beyond generic responses.

3.2.2. Retrieval-Augmented Generation (RAG)

RAG is implemented to enhance the LLM's ability to access and integrate up-to-date and domain-specific ethical knowledge, mitigating issues of hallucination and outdated information [11]. The RAG pipeline in EDA involves:

- Knowledge Base Construction: A comprehensive knowledge base is built from academic papers on design ethics, AI ethics guidelines (e.g., IEEE Ethically Aligned Design), philosophical texts, and curated ethical case studies. These documents are chunked into smaller, semantically meaningful units.
- Vector Database Indexing: Each chunk is embedded into a high-dimensional vector space using a robust embedding model (e.g., Sentence-BERT). These vectors are then indexed in a vector database for efficient semantic search.
- Retrieval Mechanism: When a designer submits a query, the system first retrieves the most relevant ethical documents or case studies from the vector database based on semantic similarity to the query. This ensures that the LLM is grounded in specific, authoritative ethical contexts.
- Augmented Generation: The retrieved documents are then provided as context to the LLM alongside the original prompt. The LLM synthesizes this retrieved information with its internal knowledge to generate a more informed and contextually relevant ethical analysis. This process can be formally represented as:

$$\text{Output}=\text{LLM}\big(\text{Prompt}+\text{Retrieve}(\text{Query},\text{KnowledgeBase})\big)$$

where Retrieve is the function that fetches relevant documents from the KnowledgeBase based on the Query.

3.2.3. Fine-Tuning

While RAG provides external knowledge, fine-tuning adapts the LLM's internal parameters to better align with the nuances of ethical reasoning in design and to improve performance on specific ethical tasks. A custom dataset of design ethics scenarios, expert ethical analyses, and corresponding ethical judgments is used for fine-tuning. This dataset includes examples of applying various ethical frameworks (e.g., Deontology, Utilitarianism, Virtue Ethics, Principlism, Capabilities Approach) to design problems. The fine-tuning process optimizes the LLM to:

- Improve Ethical Framework Application: Enhance the model's ability to consistently apply specific ethical frameworks to diverse design scenarios.
- Refine Ethical Judgment: Align the model's ethical recommendations more closely with expert human judgments.

- Reduce Bias: Mitigate algorithmic biases by exposing the model to a balanced and diverse set of ethical cases during fine-tuning.

The fine-tuning objective function can be defined as minimizing the cross-entropy loss between the model's predicted ethical analysis and the expert-annotated ground truth:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}(\theta)\right) \quad (1)$$

where θ represents the model parameters, N is the number of training examples, M is the number of possible ethical judgments/analyses, y_{ij} is 1 if the j-th judgment is the ground truth for example i and 0 otherwise, and $p_{ij}(\theta)$ is the model's predicted probability for the j -th judgment for example i.

3.3. Integration of Ethical Frameworks

The EDA system is designed to explicitly integrate and apply multiple philosophical ethical frameworks, allowing designers to analyze problems from diverse perspectives. The five primary frameworks supported are:

- Deontology: Focuses on duties and rules, assessing actions based on inherent rightness or wrongness, irrespective of consequences. The system identifies rules, duties, and whether actions adhere to them.
- Utilitarianism: Evaluates actions based on their outcomes, aiming to maximize overall well-being or happiness for the greatest number of stakeholders. The system quantifies positive and negative impacts on various stakeholders.
- Virtue Ethics: Emphasizes the character of the designer and the virtues embodied in the design process and outcome. The system prompts reflection on virtues like honesty, responsibility, and empathy.
- Principlism: Applies mid-level principles such as autonomy, beneficence, non-maleficence, and justice. The system analyzes design choices against these principles.
- Capabilities Approach: Focuses on whether a design expands or restricts individuals' capabilities and opportunities to achieve what they value. The system assesses the impact on fundamental human capabilities.

Each framework is operationalized through specific prompt engineering strategies and reinforced by fine-tuning on relevant ethical case studies, enabling the EDA to provide nuanced, multi-faceted ethical analyses.

4. Experiments and Results

To quantitatively evaluate the effectiveness and engineering value of the Ethical Deliberation Assistant (EDA) system, three complementary experiments were conducted. These experiments focused on ethical decision quality, user satisfaction, and the system's ability to apply diverse ethical frameworks. Data was collected and analyzed using statistical methods to ensure scientific rigor and verifiability.

4.1. Experimental Design

4.1.1. Experiment 1: Ethical Decision Quality Comparison

Objective: To assess the accuracy, consistency, and comprehensiveness of EDA's ethical analyses compared to human expert judgments.

Methodology: A dataset of 50 standardized design ethics case studies was compiled. Each case study presented a realistic design dilemma requiring ethical deliberation. A panel of 5 human ethics experts independently analyzed each case, providing consensus judgments on ethical dilemmas, stakeholder impacts, and proposed solutions. The EDA system then analyzed the same 50 cases. Outputs from both the EDA and human experts were rated by a separate group of 3 independent evaluators on a 5-point Likert scale (1=Poor, 5=Excellent) across four metrics: comprehensiveness, coherence, ethical soundness, and consistency. Additionally, the time taken for analysis by both the EDA and human experts was recorded.

Metrics: Accuracy: Percentage agreement between EDA's ethical judgments and expert consensus. Comprehensiveness (M, SD): How thoroughly the analysis covers all relevant ethical aspects. Coherence (M, SD): Logical flow and clarity of the ethical reasoning. Ethical Soundness (M, SD): Alignment with established ethical principles and frameworks. Consistency (M, SD): Reproducibility of ethical judgments across similar cases. Timeliness (M, SD): Time taken to generate an ethical analysis.

4.1.2. Experiment 2: User Satisfaction and Trust Metrics

Objective: To evaluate the usability, utility, and trustworthiness of the EDA system from the perspective of professional designers.

Methodology: A user study was conducted with 30 professional designers. Participants were given a series of design tasks with embedded ethical dilemmas and instructed to use the EDA system to assist their decision-making. After using the system for a predefined period, participants completed the System Usability Scale (SUS) questionnaire and a custom survey assessing usefulness, trust in recommendations, and likelihood to adopt. Semi-structured interviews were also conducted to gather qualitative feedback.

Metrics: System Usability Scale (SUS) Score (M, SD): A standardized measure of perceived usability. Usefulness (M, SD): Perceived value of the system in aiding ethical decision-making (5-point Likert). Trust in Recommendations (M, SD): Confidence in the ethical guidance provided by the EDA (5-point Likert). Likelihood to Adopt (M, SD): Willingness to integrate the EDA into their regular design workflow (5-point Likert). Overall Satisfaction (M, SD): General satisfaction with the system (5-point Likert).

4.1.3. Experiment 3: Ethical Framework Application Efficacy

Objective: To quantify the EDA system's ability to accurately and consistently apply distinct philosophical ethical frameworks.

Methodology: A specialized dataset of 50 ethical scenarios was created, with 10 scenarios specifically designed for each of the five ethical frameworks (Deontology, Utilitarianism, Virtue Ethics, Principlism, Capabilities Approach). For each scenario, the EDA was prompted to analyze it using a specific framework. The

outputs were then evaluated by human ethics experts for accuracy (whether the correct framework was applied and its principles correctly interpreted) and consistency (whether similar scenarios yielded similar framework applications).

Metrics: Framework Application Accuracy: Percentage of correct framework applications and interpretations. Framework Application Consistency: Measure of agreement in framework application across similar scenarios.

4.2. Results

4.2.1. Experiment 1: Ethical Decision Quality Comparison

Table 1 summarizes the performance of the EDA system across key ethical decision quality metrics, compared to human expert benchmarks.

TABLE I. ETHICAL DECISION QUALITY COMPARISON (N=50 CASE STUDIES, 5-POINT LIKERT SCALE FOR QUALITY METRICS)

Metric	EDA System (Mean ± SD)	Human Expert (Mean ± SD)	Improvement (%)
Comprehensiveness	4.2 ± 0.45	4.5 ± 0.30	-6.67%
Coherence	4.1 ± 0.38	4.3 ± 0.25	-4.65%
Ethical Soundness	3.9 ± 0.52	4.6 ± 0.20	-15.22%
Consistency	4.3 ± 0.41	4.0 ± 0.35	+7.50%
Timeliness (minutes)	5.2 ± 1.8	55.0 ± 12.5	+90.55%

The EDA system achieved an overall accuracy rate of 86% when compared to the expert panel's consensus judgments. This accuracy varied with case complexity, ranging from 92% on straightforward cases to 78% on highly ambiguous scenarios. Notably, the system demonstrated superior consistency (+7.50%) and significantly reduced analysis time (+90.55%), generating analyses in an average of 5.2 minutes compared to 55.0 minutes for human experts. While human experts still outperformed the EDA in comprehensiveness, coherence, and ethical soundness, the system's performance is highly competitive, especially considering the drastic reduction in time.

4.2.2. Experiment 2: User Satisfaction and Trust Metrics

The average System Usability Scale (SUS) score was 78.5 (SD=8.2), indicating a "good" range of usability. Participants rated the system highly on ease of use (M=4.1, SD=0.6) and usefulness (M=4.3, SD=0.5). Trust in the system's recommendations was moderate (M=3.7, SD=0.8), and likelihood to adopt was strong (M=4.0, SD=0.7). Overall satisfaction was high (M=4.2, SD=0.5). Qualitative interviews revealed that designers appreciated the structured approach to ethical analysis, noting that it helped them think through problems more systematically.

Table 2 presents a detailed breakdown of user satisfaction and trust metrics.

TABLE II. USER SATISFACTION AND TRUST METRICS (N=30 PROFESSIONAL DESIGNERS)

Metric	Mean (M)	Standard Deviation (SD)
System Usability Scale (SUS)	78.5	8.2
Ease of Use (1-5 Likert)	4.1	0.6
Usefulness (1-5 Likert)	4.3	0.5
Trust in Recommendations (1-5 Likert)	3.7	0.8
Likelihood to Adopt (1-5 Likert)	4.0	0.7
Overall Satisfaction (1-5 Likert)	4.2	0.5

4.2.3. Experiment 3: Ethical Framework Application Efficacy

Table 3 summarizes the EDA system's performance in applying distinct ethical frameworks.

TABLE III. ETHICAL FRAMEWORK APPLICATION EFFICACY (N=50 CASES, 10 PER FRAMEWORK)

Ethical Framework	Accuracy	Consistency
Deontology	0.92	0.89
Utilitarianism	0.88	0.86
Virtue Ethics	0.85	0.83
Principlism	0.87	0.85
Capabilities Approach	0.90	0.87

The EDA system demonstrated strong capability in applying distinct ethical frameworks. Deontology achieved the highest accuracy (0.92) and consistency (0.89), likely due to its rule-based nature aligning well with the structured, logical reasoning of language models. Utilitarianism achieved an accuracy of 0.88 and consistency of 0.86. Virtue Ethics and the Capabilities Approach showed accuracy of 0.85 and 0.90 respectively, with consistencies of 0.83 and 0.87. Principlism achieved 0.87 accuracy and 0.85 consistency. These results indicate that the EDA system can reliably apply diverse philosophical frameworks, suggesting its potential as a tool for exploring ethical problems from multiple perspectives and providing robust, framework-specific guidance.

5. ANALYSIS AND DISCUSSION

The empirical findings from our experiments provide substantial evidence supporting the core hypothesis that AI systems, specifically the Ethical Deliberation Assistant (EDA), can offer valuable and quantifiable support for ethical decision-making in engineering design. The results underscore the system's technical efficacy, practical utility, and the engineering implications of integrating advanced LLM capabilities into complex ethical reasoning processes.

5.1. Results Rationality Analysis and Technical Effectiveness Validation

The strong performance of the EDA system in Experiment 1, with an overall accuracy rate of 86% against expert judgments, validates its technical effectiveness in ethical assessment. While human experts maintained a slight edge in comprehensiveness and ethical soundness, the EDA's superior consistency (+7.50%) and remarkable timeliness (90.55% reduction in analysis time) highlight its significant engineering value. The ability to generate ethical analyses in an average of 5.2 minutes compared to 55.0 minutes for human experts represents a substantial efficiency gain, which is critical in fast-paced design and development cycles. This efficiency does not come at the cost of accuracy, as evidenced by the high agreement with expert opinions, particularly in straightforward cases (92% accuracy).

The varying accuracy with case complexity (from 92% to 78%) suggests that while LLMs can effectively process and reason over well-defined ethical scenarios, highly ambiguous or novel dilemmas still benefit from nuanced human interpretation. This finding aligns with the understanding that AI systems, despite their advanced capabilities, currently lack the lived experience and contextual understanding inherent in human moral judgment. The RAG mechanism, by grounding the LLM in a curated knowledge base, significantly contributes to reducing hallucinations and improving factual accuracy, thereby enhancing the reliability of the ethical guidance provided [15, 16]. The fine-tuning process further refines the model's ability to align with

expert ethical judgments, making its outputs more robust and contextually appropriate for design ethics [12].

5.2. Engineering Application Value

The EDA system offers several concrete engineering application benefits:

- Democratization of Ethical Expertise: Not all design teams or organizations have immediate access to dedicated ethics consultants. The EDA provides an accessible tool that can offer preliminary ethical guidance, thereby lowering the barrier to incorporating ethical considerations early in the design process. This broadens the scope of ethical review beyond specialized teams.
- Enhanced Consistency and Reproducibility: By applying standardized ethical frameworks and a systematic reasoning process, the EDA ensures a higher degree of consistency in ethical evaluations across different projects and designers. This reproducibility is a cornerstone of sound engineering practice, allowing for more reliable and comparable ethical assessments.
- Facilitation of Reflective Design Practice: The structured approach of the EDA encourages designers to systematically identify ethical dilemmas, analyze stakeholder impacts, and evaluate design choices against explicit ethical principles. This fosters a more reflective and deliberate design process, moving away from ad-hoc ethical considerations.
- Efficiency in Ethical Review: The drastic reduction in analysis time (over 90%) allows for more frequent and iterative ethical reviews throughout the design lifecycle, enabling early detection and mitigation of potential ethical issues, which can save significant resources in later development stages.
- Multi-Perspective Analysis: The system's demonstrated efficacy in applying diverse ethical frameworks (Table 3) empowers designers to explore ethical problems from multiple philosophical viewpoints. This multi-faceted analysis can lead to more nuanced and defensible ethical positions, crucial for complex socio-technical systems.

5.3. Limitations and Future Improvement Directions

Despite its promising performance, the EDA system has certain limitations that warrant further research and development. The moderate trust scores (M=3.7) in the user study indicate that while designers find the system useful, they maintain appropriate skepticism regarding AI-generated ethical guidance. This highlights the irreplaceable role of human moral agency and the need for explainable AI (XAI) features to increase transparency and build greater trust. Future work will focus on developing more robust XAI components that can articulate the reasoning process behind ethical recommendations, including the specific ethical principles applied, the data sources consulted, and the potential trade-offs considered.

Another limitation is the reliance on predefined ethical frameworks. While comprehensive, these frameworks may not capture all emergent ethical considerations in rapidly evolving technological landscapes. Future improvements will explore adaptive learning mechanisms that allow the EDA to identify novel ethical dilemmas and integrate new ethical principles as they emerge. This could involve continuous learning from new ethical case studies and expert feedback. Furthermore, while the current system focuses on textual analysis, integrating multimodal data (e.g., design mockups, user interaction data) could provide a richer context for ethical deliberation.

Finally, the generalizability of the experimental results could be enhanced by expanding the dataset of case studies to cover a broader range of design domains and ethical complexities. Future experiments will also involve a larger and more diverse cohort of professional designers to further validate the system's usability and impact in various organizational settings. Investigating the long-term impact of EDA on design outcomes and organizational ethical culture would also be a valuable direction.

6. Conclusion

This paper presented the Ethical Deliberation Assistant (EDA), an AI-driven system designed to enhance ethical decision-making in engineering design through the technical integration of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and fine-tuning. We introduced a robust technical framework detailing the system's architecture and the application of advanced AI techniques to operationalize ethical reasoning. Our quantitative evaluation demonstrated the EDA's significant engineering value, achieving an 86% accuracy rate in ethical judgments compared to human experts and reducing analysis time by over 90%.

The EDA system provides a structured, consistent, and efficient mechanism for designers to navigate complex ethical dilemmas, democratizing access to ethical expertise and fostering a more reflective design practice. Its ability to reliably apply diverse ethical frameworks (with Deontology achieving 0.92 accuracy) further underscores its utility in multi-perspective ethical analysis. While acknowledging limitations related to human trust and the need for enhanced explainability, this work establishes a foundational technical approach for responsible AI development in design. Future research will focus on integrating advanced XAI features, adaptive learning for emergent ethical considerations, and expanding experimental validation across broader design contexts to further solidify the EDA's role as an indispensable tool for ethical engineering.

References

[1] Jiao, J., Afroogh, S., Murali, A., Chen, K., Atkinson, D., & Dhurandhar, A. (2025). LLM ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. Scientific Reports, 15(1), 34642.

[2] Kapania, S., Wang, R., Li, T. J. J., Li, T., & Shen, H. (2025). 'I'm Categorizing LLM as a Productivity Tool': Examining Ethics of LLM Use in HCI Research Practices. Proceedings of the ACM on Human-Computer Interaction, 9(2), 1-26.

[3] Singh, S. (2025). Systems Engineering of Large Language Models for Enterprise Applications.

[4] Zhou, Y., & Chen, C. H. (2025). Examining the impact of large language models on design: Functions, strengths, limitations, and roles. Design and Artificial Intelligence, 100017.

[5] Esmaeilzadeh, P. (2025). Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Medical Informatics and Decision Making, 25(1), 342.

[6] Hadar-Shoval, D., Asraf, K., Shinan-Altman, S., Elyoseph, Z., & Levkovich, I. (2024). Embedded values-like shape ethical reasoning

of large language models on primary care ethical dilemmas. Heliyon, 10(18).

[7] Ferdaus, M. M., Abdelguerfi, M., Loup, E., N. Niles, K., Pathak, K., & Sloan, S. (2026). Towards trustworthy AI: a review of ethical and robust large language models. ACM Computing Surveys, 58(7), 1-43.

[8] Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., ... & Tenenbaum, J. B. (2022). Computational ethics. Trends in cognitive sciences, 26(5), 388-405.

[9] Alkayyal, M., Malberg, S., & Groh, G. (2025, September). An LLM-Based Decision Support System for Strategic Decision-Making. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 460-464). Cham: Springer Nature Switzerland.

[10] Arif, S., Amjad, M. U., & Faisal, M. (2025). AI-Driven Decision Support Systems for Software Architecture: A Framework for Intelligent Design Decision-Making (2025). Journal of Computing and Artificial Intelligence, 3(1).

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[12] Pradhan, R. (2025). RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models. International Journal of Computer Technology and Electronics Communication, 8(5), 11326-11333.

[13] Wang, Z., Liu, Z., Lu, W., & Jia, L. (2025). Improving knowledge management in building engineering with hybrid retrieval-augmented generation framework. Journal of Building Engineering, 103, 112189.

[14] Wan, Y., Chen, Z., Liu, Y., Chen, C., & Packianather, M. (2025). Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. Advanced Engineering Informatics, 65, 103212.

[15] Gan, A., Yu, H., Zhang, K., Liu, Q., Yan, W., Huang, Z., ... & Hu, G. (2025). Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. arXiv preprint arXiv:2504.14891.

[16] Collaco, B. G., Srinivasagam, P., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Wood, N. G., ... & Forte, A. J. (2026). Integrating Fine-Tuning and Retrieval-Augmented Generation for Healthcare AI Systems: A Scoping Review. Bioengineering, 13(2), 225.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

ETHICAL STATEMENT

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

AUTHOR CONTRIBUTIONS

Lingyan Zhang conceived and designed the Ethical Deliberation Assistant (EDA) framework, developed the system architecture integrating prompt engineering, Retrieval-Augmented Generation (RAG), and fine-tuned large language models, conducted the experimental evaluation, analyzed and interpreted the results, and wrote the manuscript, while Xusheng Zhang contributed to the design of the ethical assessment methodology, assisted with data collection and statistical analysis, validated the experimental findings, provided critical revisions to the manuscript, and supervised the study and final manuscript preparation.

COMPETING INTERESTS

The authors declare no competing interests.