# Mitigating Values Debt in Generative AI: Responsible Engineering with Graph RAG

Waqar Hussain*
*CSIRO, Melbourne, Australia
Email: waqar.hussain@data61.csiro.au

2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE) | 979-8-3315-1466-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/RAIE66699.2025.00006

***Abstract*—Generative AI technologies are rapidly transforming industries such as healthcare, education, and transportation. However, this progress often incurs a Values Debt—ethical and operational deficits due to insufficient ethical considerations during development. This paper examines Values Debt in Generative AI and introduces the Helpful, Honest, Harmless (HHH) framework to align AI systems with human values. In developing GRAISE, a Graph RAG-based chatbot for aviation safety, the HHH framework is applied to integrate ethical practices throughout the development process. This case study demonstrates how the HHH framework addresses ethical challenges and provides reliable contextual information to enhance pilot communication, exemplifying responsible AI engineering. These findings advocate for the broader adoption of ethical AI frameworks across various sectors, promoting trust and integrity in AI applications.**

***Index Terms*—Generative AI, Values Debt, HHH Framework, Ethical AI, Responsible AI, AI Safety, AI Bias, Hallucination**

## I. Introduction

Generative AI (GenAI) technologies like GPT-4 and DALL.E have transformed industries by automating tasks such as content creation, language translation, and data analysis [1]. These advancements have driven efficiency and innovation in fields like healthcare, education, and transportation [**?**]. However, they also introduce critical ethical challenges, including biases, misinformation, and privacy risks [2]. Traditional software engineering practices prioritize code correctness and performance but often neglect these ethical complexities.

This paper introduces the concept of *Values Debt*—the ethical and operational deficits arising from insufficient attention to ethical principles during AI development [3]. To address this, we propose integrating the *Helpful, Honest, Harmless* (HHH) framework [4] into AI engineering practices. The HHH framework ensures AI systems are *helpful* by meeting user needs, *honest* through accurate and transparent outputs, and *harmless* by preventing societal harm and fostering ethical operation. Using a case study of a chatbot developed using Graph Retrieval Augmented Generation (RAG) [5] for aviation safety, we demonstrate how this framework effectively mitigates Values Debt and aligns AI systems with human values.

This paper highlights how Values Debt manifests in GenAI-based applications and addresses the key research question:

- How can the HHH framework mitigate Values Debt and align AI systems with human values?

## II. Understanding Values Debt

Values Debt, analogous to technical debt, refers to the ethical and operational deficits that accumulate during the development and deployment of AI systems [6]. Unlike technical debt, which arises from coding shortcuts, Values Debt stems from neglecting ethical considerations, often remaining hidden until after deployment and complicating remediation efforts.

The development of Generative AI is driven by diverse intentions, ranging from benign innovation to malicious exploitation, influenced by factors such as the pursuit of innovation, power dynamics, and market dominance [3], [6]. This rapid advancement prioritizes short-term gains over long-term ethical standards, leading to significant Values Debt that compromises the integrity and societal trust in AI technologies.

### A. Manifestations and Implications of Values Debt

Generative AI systems frequently launch with inherent biases, inaccuracies, and ethical issues due to poor-quality or non-inclusive training datasets sourced from internet wikis, websites, blogs, and videos. The rush to capture market share, achieve short-term financial gains, adhere to research agendas, and occasional oversight compromises ethical rigour, resulting in long-term societal and organizational risks [3], [6], [7].

*1) Disinformation and Malicious Applications:* Generative AI's ability to produce realistic yet false content facilitates disinformation campaigns, deepfakes, and fraud. AI hallucinations—where systems generate plausible but inaccurate information—exacerbate misinformation risks, undermining public trust and threatening security [2].

*2) Social Injustice and Bias:* AI models often replicate and amplify societal biases present in their training data, resulting in discriminatory outcomes. For example, Amazon's AI recruitment tool favoured male candidates due to biased data, leading to its discontinuation [7]. Similarly, Google's internal conflicts over ethical AI practices, highlighted by the departure of AI researcher Timnit Gebru, reveal organizational tensions from misaligned values [8].

*3) Reputational and Legal Implications:* Values Debt can lead to reputational harm, legal challenges, and loss of consumer trust. Unethical AI systems risk public backlash, lawsuits, and diminished market share [9]. Addressing these issues post-deployment is more costly and complex than integrating ethical considerations during development [3].

Addressing Values Debt requires embedding ethical considerations throughout the AI development lifecycle. By moving from a purely functional approach to one that incorporates ethical principles, AI systems can align with the *Helpful, Honest, Harmless* (HHH) framework [4].

979-8-3315-1466-2/25/$31.00 ©2025 IEEE
DOI 10.1109/RAIE66699.2025.00006

## III. Case Study: GRAISE—an Aviation Chatbot

Aviation safety traditionally relied on redundant systems and rigid procedures to mitigate risk. While effective, these measures often restrict operational flexibility and limit adaptive learning. Generative AI offers a transformative approach by enabling technologies to adapt to pilots' learning needs [10]. Addressing this, GRAISE enhances pilots' communication and decision-making under pressure—critical skills for safe operations [11].

To operationalize the *Helpful, Honest, Harmless* (HHH) framework and mitigate Values Debt, GRAISE integrates GPT-4 with a graph-structured aviation knowledge base, while adhering to HHH principles throughout system development. This integration provides adaptive, context-sensitive feedback, enhancing the realism of training scenarios and improving the effectiveness of pilot training programs.

### *A. Graph RAG Implementation*

We built a Neo4j knowledge graph using NASA's ASRS incident reports and relevant communication theories, capturing key entities (e.g., `CommunicationBreakdown`, `Aircraft`, `Incident`, `Anomaly`) and their relationships. User queries map to these graph nodes and edges, anchoring GPT-4's generative output in validated domain data. This structure mitigates misinformation risks, provides transparency about data provenance, and aligns with the Helpful, Honest, Harmless principles.

Graph RAG leverages this knowledge graph by employing community detection algorithms to group related entities into "communities." Each community is summarized in parallel, and the partial summaries are merged into a comprehensive final answer using a map-reduce strategy [5]. By structuring data graphically, Graph RAG reduces hallucinations, provides clearer causal links, and enhances both the diversity and completeness of answers—especially in large-scale or high-stakes contexts.

Empirical evaluations demonstrate that Graph RAG significantly outperforms naïve RAG in both comprehensiveness and diversity of answers, making it ideal for high-stakes applications like aviation safety that require global insights. Additionally, Graph RAG's intermediate and low-level community summaries surpass source text summarization on these metrics while reducing token costs [5].

GRAISE leverages Graph RAG to enhance pilot training by integrating GPT-4 with the aviation-specific knowledge graph. By evaluating incident narratives through established communication theories [11], GRAISE provides context-sensitive feedback that supports pilot decision-making and mitigates ethical risks tied to misinformation.

### *B. Embedding the HHH Framework in GRAISE Development*

While Graph RAG strengthens GRAISE's outputs by minimizing hallucinations, improving transparency, and grounding responses in validated knowledge, achieving truly ethical outcomes demands a holistic approach that embeds responsible practices throughout the development lifecycle. By incorporating the HHH framework at each stage—from requirements engineering and data management to deployment and ongoing monitoring—GRAISE not only provides more reliable and transparent feedback but also mitigates Values Debt. The result is an AI system whose outputs and underlying processes are aligned with fairness, accountability, and the ethical standards essential for safe and trustworthy pilot training.

**Requirements Engineering**: During this phase, ethical requirements were defined alongside functional specifications to include fairness, accountability, and transparency [12]. Engaging with airline pilots during early prototyping ensured that GRAISE's objectives—enhancing pilot training while avoiding unnecessary complexity—matched real-world stakeholder needs. This collaborative process reduced the risk of ethical deficits from the outset.

**Data Management**: Ensuring that the data used for training GRAISE was ethically sourced, diverse, and representative was a critical focus [13]. The aviation knowledge graph was constructed from a sample of 97 anonymized communication-related incident reports and safety data from the publicly available ASRS reporting tool [14], minimizing biases and ensuring meaningful coverage of communication scenarios. Data governance policies were established to oversee data quality and compliance, thereby preventing the dissemination of biased or inaccurate information. Privacy preservation measures were integrated by anonymizing data and ensuring no personal or sensitive information was compromised [14].

**Design and Architecture**: GRAISE's design prioritizes explainability, accuracy, and user control through the development of a communication-specific aviation ontology and the construction of a semantically searchable knowledge graph enriched with trusted data from NASA's incident reports [14]. Semantic rules and constraints ensure consistent and precise data representation, while optimizing external knowledge enhances retrieval efficiency and relevance. By adopting a Graph RAG framework [5], GRAISE integrates large language models with structured aviation data, enabling context-sensitive decision-making. Traceable knowledge retrieval and reasoning processes enhance transparency and foster user trust, resulting in a reliable and trustworthy AI system that effectively minimizes Values Debt.

**Collaborative Development with Key Stakeholders**: GRAISE was initially developed through low-fidelity prototyping and co-design with airline pilots, ensuring that it met the actual needs of users and addressed potential ethical concerns [2]. This collaboration enhances the helpfulness and harmlessness of GRAISE by aligning it closely with user expectations and domain requirements.

**Verification and Validation**: Developing robust testing protocols for ethical compliance was essential in the verification and validation phase. GRAISE underwent rigorous evaluation by experts and aviation pilots of diverse genders to detect and mitigate biases, ensuring that the AI outputs remain fair and accurate. These evaluations included scenario-based testing with pilots, allowing for continuous refinement of the system to better align with ethical standards.

**Deployment and Monitoring**: Although GRAISE is still a prototype, it is being tested in controlled environments where domain experts, pilots, and other stakeholders continuously evaluate its performance and ethical integrity. Potential misinformation tied to GPT-4's broader training data is monitored, and safeguards help detect misuse. Feedback loops allow iterative updates, preserving GRAISE's responsible AI objectives as it evolves toward operational readiness. Ongoing monitoring thus preserves GRAISE's responsible AI objectives throughout its operational lifecycle.

## IV. Alignment of GRAISE with the HHH Framework

GRAISE exemplifies the *Helpful, Honest, Harmless* (HHH) framework, effectively mitigating Values Debt in a safety-critical context. The following outlines how GRAISE aligns with each HHH principle:

**Helpful:** GRAISE assists pilots by delivering personalized feedback from real-world incident reports, enhancing communication and decision-making skills. Utilizing a knowledge graph developed with an aviation-specific ontology, it provides accurate, contextually relevant responses tailored to pilots' needs to understand communication safety. By analyzing ASRS safety narratives, GRAISE offers concise, actionable insights that directly support pilot training objectives.

**Honest:** GRAISE anchors its outputs in verified aviation safety data, minimizing the hallucinations common in large language models (LLMs). By integrating a knowledge graph based on the ASRS taxonomy and communication theories, GRAISE ensures transparency and verifiability. The system clearly communicates its capabilities and limitations by informing users when information is unavailable in the knowledge base, rather than fabricating responses or relying on GPT-4's inherent training data. This honest approach fosters user trust and enhances the reliability of the AI system.

**Harmless:** By improving pilots' ability to manage communication breakdowns and interpersonal challenges, GRAISE helps reduce aviation accident risks. It avoids harmful or misleading advice through reliance on validated data and adherence to safety protocols. GRAISE ensures respectful and professional interactions, preventing offensive or discriminatory content.

By integrating LLMs with domain-specific knowledge, GRAISE enhances safety-critical training while addressing Values Debt. Delivering accurate, safe, and data-driven feedback, GRAISE exemplifies the HHH framework's real-world impact, fostering trust and safety in aviation.

## V. Implications for Other Domains

The successful implementation of the HHH framework using a Graph RAG approach in aviation offers valuable insights for other fields seeking to enhance the ethical dimensions of their AI applications. With this approach, organizations can address common ethical challenges such as misinformation, lack of transparency, and unintended biases. This section examines how the observed benefits can enhance functionality and ethical compliance across various AI domains.

### *A. Enhancing Accuracy through Structured Knowledge*

Applying Graph RAG allows AI systems to ground their outputs in verified, structured data pertinent to a specific domain. This grounding can significantly reduce instances of hallucinations and misinformation prevalent in conventional LLMs. For example, in the legal sector, integrating statutes and case law into a legal knowledge graph could enable AI assistants to provide accurate legal research and analysis, enhancing the helpfulness of the system while maintaining factual correctness.

### *B. Improving Transparency and Explainability*

The use of knowledge graphs facilitates the traceability of information, as the relationships and sources within the graph can be inspected and audited. This transparency aligns with the honesty principle of the HHH framework. In domains like finance, where understanding the rationale behind investment advice is crucial, AI systems that can reference economic indicators and market data from a financial knowledge graph would offer more explainable and trustworthy support to users.

### *C. Mitigating Bias and Promoting Fairness*

By carefully curating the data included in a knowledge graph and continuously monitoring for biases, AI applications can better adhere to the harmlessness principle. In the field of education, ensuring that the knowledge graph includes diverse curricular content can help AI tutors provide equitable learning experiences, promoting fairness in educational guidance.

### *D. Enhancing Data Privacy and Security*

Graph RAG frameworks can be designed to handle sensitive information securely, as demonstrated with de-identified incident reports in aviation. In customer service, AI chatbots utilizing product knowledge graphs can offer personalized assistance while safeguarding customer data, aligning with privacy regulations and ethical standards.

### *E. Facilitating Adaptive and Personalized Interactions*

The adaptability of Graph RAG systems enables AI applications to provide context-sensitive and personalized responses. This adaptability enhances the helpfulness of the AI and supports users in complex decision-making processes. For example, in environmental conservation, AI systems grounded in ecological knowledge graphs can provide tailored recommendations for sustainable practices, benefiting both users and the environment.

### *F. Supporting Ethical AI Development Practices*

The integration of the HHH framework with Graph RAG encourages developers to consider ethical principles from the outset of AI system design. This proactive approach can lead to the creation of AI applications that are not only effective but also aligned with societal values and regulatory requirements. By adopting these practices, organizations across various domains can contribute to building public trust in AI technologies.

Aviation, as an example industry, is increasingly leveraging Generative AI to enhance predictive maintenance, workflow automation, record management, and customer service [15], [16]. Leading organizations such as Airbus, Boeing, GE Aerospace, and Manchester Airports Group utilize AI for real-time analytics, AI-assisted maintenance, and improving passenger experiences [15], [16].

## VI. Discussion

Integrating ethical considerations into AI development requires a substantial cultural shift in software engineering practices. The GRAISE chatbot exemplifies how operationalizing the HHH framework can mitigate Values Debt in high-stakes environments such as aviation, providing valuable insights that can be applied to other sectors.

**Balancing Technical Feasibility with Ethics:** Implementing HHH principles may demand extra resources and could affect system performance [17]. Nevertheless, GRAISE's experience shows that ethical diligence yields long-term benefits in trust and compliance. Weighing these factors early in the design phase ensures that ethical imperatives remain central, even under resource constraints.

**Continuous Evaluation and Adaptation:** As ethical standards and domain regulations evolve, development processes must remain adaptable. Approaches like Graph RAG enable the integration of updated knowledge and standards, facilitating ongoing evaluation of AI systems. This flexibility highlights the need for AI architectures that can swiftly respond to shifting ethical and regulatory requirements.

**Cross-Disciplinary Collaboration:** GRAISE's development underscored the value of engaging technologists, domain experts, ethicists, and stakeholders [18]. Such collaboration helps identify and mitigate ethical risks proactively, leading to technically robust and socially responsible systems.

**Privacy, Safety, and Broader AI Implications:** Protecting user data and ensuring safety are essential to the harmlessness principle. Implementing data anonymization, secure access protocols, and grounding AI in verified information helps meet regulations like GDPR and builds user trust by preventing harmful outputs. These approaches provide a blueprint for AI applications in other domains, where structured knowledge, bias mitigation, and transparent decision-making can enhance ethical performance.

## VII. Conclusion

Values Debt—ethical and operational deficits arising from insufficient ethical considerations during development—poses significant risks in generative AI application development. This study addresses Values Debt by operationalizing the core ethical principles of being Helpful, Honest, and Harmless (HHH) through ethical system engineering practices and technical approaches such as Graph RAG. This strategy ensures the integration of domain knowledge, transparency of information, and application reliability, thereby enhancing user trust.

The GRAISE case study exemplifies the practical application of these principles and methodologies. By leveraging Graph RAG, GRAISE demonstrates the creation of ethically sound and socially beneficial AI, highlighting the effectiveness of the proposed strategies across various domains.

Through proactive governance and careful stewardship, the development of generative AI can enhance societal well-being while maintaining ethical integrity. This commitment fosters a symbiotic relationship where humans and AI collaboratively shape each other for mutual benefit.

## Acknowledgment

The author gratefully acknowledges Dr. Fabilo Mattioli for his insightful discussions on the GRAISE project.

## References

[1] OpenAI *et al.*, "Gpt-4 technical report," OpenAI, Tech. Rep., 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[2] L. Weidinger *et al.*, "Taxonomy of ethical and social risks of harm from language models," *arXiv preprint arXiv:2212.08073*, 2022.

[3] W. Hussain, "Values debt is eating software," *IEEE Software Blog*, vol. 12, 2019. [Online]. Available: https://software.ieee.org/blog/values-debt-is-eating-software

[4] A. Askell *et al.*, "A general language assistant as a laboratory for alignment," *arXiv preprint arXiv:2112.00861*, 2021.

[5] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024. [Online]. Available: https://arxiv.org/abs/2404.16130

[6] W. Hussain, "Thorns and algorithms: Navigating generative ai challenges inspired by giraffes and acacias," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2407.11360

[7] R. Iriondo, "Amazon scraps secret ai recruiting engine that showed biases against women," 2018. [Online]. Available: https://shorturl.at/IvvbQ

[8] D. Wakabayashi and C. Metz, "Another firing among google's a.i. brain trust, and more discord," *The New York Times*, May 2022, accessed: 2024-11-19. [Online]. Available: https://shorturl.at/WefK0

[9] T. Walsh, *Machines That Think: The Future of Artificial Intelligence*. Prometheus Books, 2018.

[10] Federal Aviation Administration, "Leadership and command training for pilots in command," 2020. [Online]. Available: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_121-42.pdf

[11] K. Perkins, S. Ghosh, J. Vera, C. Aragon, and A. Hyland, "The persistence of safety silence: How flight deck microcultures influence the efficacy of crew resource management," *International Journal of Aviation, Aeronautics, and Aerospace*, vol. 9, no. 3, 2022. [Online]. Available: https://doi.org/10.15394/ijaaa.2022.1728

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

[13] C. A. Bail *et al.*, "Can we mitigate bias in ai?" *Annual Review of Sociology*, 2023.

[14] NASA Aviation Safety Reporting System, "Asrs database online," https://asrs.arc.nasa.gov/search/database.html, 2022, accessed: 2024-11-19.

[15] GEAerospace, "Ge aerospace teams up with microsoft and accenture to unveil generative ai-powered solution for faster, more informed aircraft service and maintenance record insights," 2024. [Online]. Available: https://shorturl.at/K46LB

[16] N. Woods and R. Cant, "Manchester airports group looks to AWS to transform the passenger experience," 2024. [Online]. Available: https://shorturl.at/YTsFW

[17] W. Hussain, M. Shahin, R. Hoda, J. Whittle, H. Perera, A. Nurwidyantoro, R. A. Shams, and G. Oliver, "How can human values be addressed in agile methods? a case study on safe," *IEEE Transactions on Software Engineering*, vol. 48, no. 12, pp. 5158–5175, 2022.

[18] B. D. Mittelstadt *et al.*, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, 2016.