A Trustworthy Generative AI Risk Assessment Framework for Enterprise Information Systems

Donggao Shao
Department of Computer Science, University of North Texas, Denton, TX, USA.
donggaowork@unt.edu

Deepak Saini
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
hellodeepak@uc.edu

Dev D. Batra
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
dev.d.batra@uab.edu

Abstract

The rapid integration of generative artificial intelligence into enterprise information systems presents unprecedented opportunities for automation, personalization, and knowledge synthesis, yet simultaneously introduces complex risks that challenge established governance and security paradigms. Existing risk assessment frameworks often fail to account for the unique properties of generative models, including emergent capabilities, opaque reasoning paths, and susceptibility to adversarial manipulation. This paper proposes a comprehensive risk assessment framework designed specifically for trustworthy generative AI within enterprise contexts, emphasizing systemic evaluation across technical, organizational, and regulatory dimensions. The framework integrates principles from socio-technical systems theory, software engineering, and responsible AI to address structural trade-offs between innovation and control, performance and interpretability, and autonomy and oversight. Architecture-level considerations such as model provenance, data lineage, and deployment topology are examined alongside governance mechanisms including continuous monitoring, audit trails, and incident response protocols. Special attention is given to fairness, robustness, and sustainability as cross-cutting concerns that must be embedded throughout the lifecycle of generative AI services. The framework further incorporates policy implications by mapping compliance requirements from emerging regulations onto concrete risk metrics and decision thresholds. Through detailed analytical discussion and illustrative case comparisons, the paper demonstrates how enterprises can systematically evaluate generative AI deployments without stifling beneficial use cases. The proposed approach moves beyond checklist-based assessment toward a dynamic, context-aware methodology that evolves with model updates and shifting operational environments. By aligning risk management with trustworthiness principles, the framework provides a foundation for enterprises to responsibly harness generative AI while maintaining accountability and resilience in their information systems.

Keywords

generative AI, risk assessment, enterprise information systems, trustworthiness, governance, socio-technical systems, fairness, robustness, sustainability, regulatory compliance.

1. Introduction

Generative artificial intelligence has transitioned from experimental research to widespread enterprise adoption, powering applications that range from automated content creation and code generation to conversational agents and predictive analytics. Unlike traditional discriminative models, generative models produce novel outputs that can exhibit unexpected behaviors, making their integration into mission-critical information systems a matter of considerable concern. The very characteristics that make generative AI powerful, such as its ability to synthesize plausible information and adapt to diverse contexts, also introduce vulnerabilities that conventional risk assessment methods are ill-equipped to handle. Enterprises seeking to deploy generative AI must navigate a landscape where model failures can cascade through interconnected systems, propagate biased or harmful content, and expose sensitive data through memorization or inference attacks. The challenge is compounded by the rapid pace of model development, the opacity of large-scale architectures, and the evolving regulatory environment.

Existing risk assessment frameworks in information security and software engineering, such as those based on ISO 31000 or NIST SP 800-30, provide general guidelines but lack specificity for generative AI. These frameworks typically assume deterministic systems with well-defined failure modes, whereas generative models operate in a probabilistic space where outputs are conditioned on vast, often uncurated training data. Moreover, the emergent capabilities of large language models and diffusion models create risks that cannot be anticipated from component-level analysis alone. Therefore, a dedicated framework is needed to address the unique interplay between technical reliability, organizational readiness, and societal expectations. This paper responds to that need by proposing a trustworthy generative AI risk assessment framework that is both comprehensive and adaptable to enterprise contexts.

The framework is grounded in systems thinking, recognizing that generative AI does not operate in isolation. It is embedded within a socio-technical infrastructure consisting of data pipelines, human oversight teams, application interfaces, and external regulatory constraints. Any risk assessment must therefore consider not only the model itself but also the surrounding processes, incentives, and feedback loops. The framework draws on prior work in AI safety, fairness, and transparency [1,2,3], as well as established enterprise risk management principles [4,5]. By synthesizing these strands, the framework aims to provide a structured yet flexible methodology that can be tailored to different organizational scales and industry sectors.

The remainder of the paper is organized as follows. Section 2 reviews related work and situates the proposed framework within the broader literature. Section 3 describes the architecture and core components of the framework. Section 4 examines the key dimensions of trustworthiness that the framework operationalizes. Section 5 details the risk assessment methodology, including quantitative and qualitative metrics. Section 6 addresses governance and policy implications, including compliance with emerging regulations. Section 7 discusses deployment and sustainability considerations, including resource consumption and model lifecycle management. Section 8 concludes with a synthesis of findings and directions for future work.

2. Background and Related Work

The landscape of AI risk assessment has evolved considerably over the past decade, driven by high-profile incidents involving biased algorithms, privacy breaches, and system failures. Early efforts focused on fairness metrics for classification systems [6] and adversarial robustness for deep neural networks [7]. These works provided foundational tools but were

developed primarily for discriminative models and static datasets. With the advent of generative models, researchers began to explore risks specific to language generation, such as toxic output, hallucination, and data contamination [8,9]. The concept of AI trustworthiness emerged as a broader umbrella encompassing not only accuracy and safety but also explainability, accountability, and value alignment [10]. This shift reflects a growing recognition that technical performance alone is insufficient for enterprise deployment; organizational and societal factors must be integrated into assessment.

Enterprise information systems have long employed risk management frameworks such as COBIT, ISO 27001, and the FAIR model. These frameworks are designed to handle threats like unauthorized access, data loss, and system downtime, but they do not natively address the stochastic and generative nature of AI outputs. Several recent proposals have attempted to extend these frameworks to AI, for instance by adding model risk management layers [11] or by adopting the concept of machine learning operations (MLOps) to formalize monitoring and governance [12]. Yet, these extensions often treat generative AI as a special case of machine learning, overlooking fundamental differences in how risk manifests. For example, a generative model can produce harmful content without any explicit malicious input, a scenario not covered by traditional threat models.

The proposed framework builds upon these prior efforts but takes a more holistic view by embedding trustworthiness principles directly into the risk assessment process. Rather than adding a separate module for AI risk, the framework redefines the risk evaluation pipeline to account for generative capabilities from the outset. This aligns with recent calls for responsible AI governance that integrates technical, organizational, and ethical considerations [13]. The framework also incorporates insights from socio-technical systems theory, which emphasizes that the reliability of complex systems depends on the interaction between human operators and automated components [14]. In enterprise settings, generative AI is often used in human-in-the-loop configurations, making socio-technical factors critical to risk assessment.

3. Framework Architecture and Components

The proposed risk assessment framework is structured around three interconnected layers: the technical layer, the organizational layer, and the regulatory layer. Each layer contains a set of components that together enable a comprehensive evaluation of generative AI risk within the enterprise. The technical layer addresses model behavior, data quality, and infrastructure vulnerabilities. The organizational layer focuses on human oversight, decision processes, and incident response. The regulatory layer ensures alignment with current and anticipated legal requirements. The layers are not hierarchical but interact dynamically; for example, a technical vulnerability may require organizational changes to mitigate, and regulatory compliance may place constraints on technical design choices.

At the core of the technical layer is the concept of model provenance. Unlike traditional software, where source code is fully controlled, generative models are often derived from pre-trained checkpoints that may contain undocumented biases or vulnerabilities. Assessing provenance involves tracking the training data, architecture, fine-tuning procedures, and any modifications applied during deployment [15]. This lineage information is essential for understanding the model's inherent risks, such as its tendency to reproduce toxic content or its susceptibility to adversarial prompt injection. The framework mandates that every generative AI component in an enterprise system have a documented provenance record that is auditable and version-controlled.

Complementing provenance is the data lineage component, which maps the flow of data from collection through preprocessing to model training and inference. Data quality issues such as labeling errors, skew, and omission can lead to unreliable outputs. In generative contexts, training data may include copyrighted material, personal identifiers, or hate speech, raising legal and ethical concerns. The framework requires enterprises to maintain a data inventory that specifies the sources, licenses, and any privacy protections applied. This inventory serves as the basis for risk scoring related to data rights violations and model fidelity.

The organizational layer addresses the human and procedural dimensions that mediate risk. Enterprises must define clear roles and responsibilities for generative AI oversight, including a model risk owner, a compliance officer, and a technical lead. The framework recommends establishing a cross-functional review board that evaluates each deployment against predefined trustworthiness criteria before release. This board should have the authority to halt deployment if risk thresholds are exceeded. Additionally, incident response protocols must be tailored to generative AI failures, which may require rollback to a previous model version, output filtering, or notification of affected users. The framework emphasizes that organizational readiness is as important as technical robustness; a highly accurate model can still cause harm if deployed without adequate safeguards.

The regulatory layer aligns the framework with external requirements, such as the European Union's AI Act, the proposed U.S. AI Bill of Rights, and sector-specific regulations in finance, healthcare, and law. These regulations often mandate transparency, documentation, and human oversight. The framework maps these requirements to specific technical and organizational controls, enabling enterprises to present a compliance case to auditors. For instance, the requirement for explainability can be met by implementing interpretability tools such as influence functions or attention visualization, while the requirement for non-discrimination can be satisfied by conducting fairness audits on benchmark datasets [16].

4. Trustworthiness Dimensions

Trustworthiness in generative AI extends beyond traditional reliability to encompass several interrelated dimensions: accuracy, fairness, robustness, interpretability, privacy, accountability, and sustainability. Each dimension must be assessed within the specific context of the enterprise application, as trade-offs are inevitable. For example, increasing interpretability may reduce model performance, and enhancing robustness against adversarial inputs may require additional computational resources that affect sustainability. The framework does not prescribe a single optimal configuration but provides tools to evaluate these trade-offs explicitly and transparently.

Fairness is a particularly challenging dimension for generative AI because models can produce outputs that perpetuate or amplify societal biases even when training data appears balanced. The framework operationalizes fairness through multiple metrics, including demographic parity, equal opportunity, and counterfactual fairness, adapted for generative tasks such as text generation and image synthesis [17]. These metrics are computed not only on test sets but also on live deployments through sampling and user feedback. Enterprises must define acceptable deviation thresholds and implement corrective mechanisms, such as debiasing fine-tuning or output reranking, when thresholds are breached.

Robustness addresses the model's resistance to both accidental inputs, such as typos or ambiguous queries, and malicious attacks, such as prompt injection or data poisoning. Generative models are particularly vulnerable to adversarial examples that cause the model to

produce harmful or misleading content. The framework recommends stress-testing models using a library of adversarial scenarios tailored to the deployment domain. For enterprise systems handling financial transactions or medical advice, robustness failures can have severe consequences, so the framework requires that any model with high failure impact undergo additional validation, including red-teaming exercises [18].

Accountability refers to the ability to attribute outcomes to specific decisions, model versions, and actors. The framework mandates logging all inference requests, outputs, and human interventions in a tamper-evident manner. These logs enable post-hoc analysis if an incident occurs, allowing the organization to identify root causes and implement corrective measures. Accountability also requires that the enterprise designate a responsible entity for each deployment, ensuring that there is a clear point of contact for external stakeholders such as regulators or affected users.

Sustainability has emerged as a critical concern given the enormous computational resources required to train and serve large generative models. The framework incorporates a sustainability dimension that evaluates energy consumption, carbon footprint, and hardware utilization. Enterprises are encouraged to consider model compression, efficient inference techniques, and renewable energy sources as part of their risk mitigation strategy [19]. While sustainability is not always viewed as a direct risk, societal pressure and potential future regulations make it prudent to include it in the assessment.

5. Risk Assessment Methodology

The methodology follows a structured process comprising five phases: scoping, identification, analysis, evaluation, and treatment. In the scoping phase, the enterprise defines the boundaries of the generative AI system, including the models, data, interfaces, and human operators involved. This phase also establishes the risk appetite of the organization and the acceptable thresholds for each trustworthiness dimension. The identification phase systematically enumerates potential failure modes using techniques such as hazard analysis and scenario brainstorming. For generative AI, typical hazards include output toxicity, factuality errors, privacy leakage, and adversarial manipulation. Each hazard is characterized by its cause, consequence, and potential severity.

The analysis phase assigns likelihood and impact scores to each identified risk. Likelihood is estimated based on historical data from similar deployments, empirical testing, and expert judgment. Impact is assessed in terms of financial loss, reputational damage, legal liability, and harm to individuals. The framework uses a multi-attribute approach that allows combining these factors into a composite risk score. Importantly, the analysis accounts for interdependencies between risks; for example, a fairness violation may amplify privacy risks if biased outputs disclose sensitive information about underrepresented groups. The evaluation phase compares the risk scores against the predefined thresholds, categorizing risks as acceptable, tolerable, or unacceptable. Risks in the tolerable region require documented mitigation plans, while unacceptable risks mandate a stop to deployment.

The treatment phase selects and implements measures to reduce risk to acceptable levels. These measures can be technical, such as adding an output filter to block toxic content, or organizational, such as augmenting human oversight during peak usage periods. The framework emphasizes that treatment should be proportionate to the risk and that residual risk must be accepted by the responsible decision-maker. An iterative loop ensures that after treatment, the risk is re-evaluated to confirm that it falls within acceptable bounds.

Throughout this process, the framework encourages the use of continuous monitoring to detect drift in model behavior or changes in the external environment that could alter risk levels [20].

6. Governance and Policy Implications

Governance structures for generative AI must be embedded within the enterprise's broader risk management framework. The proposed model advocates for a three-tier oversight system: operational oversight at the team level, tactical oversight at the departmental level, and strategic oversight at the executive level. This tiered approach ensures that risk decisions are made with appropriate context and authority. For instance, a product team may decide to accept a moderate risk of harmless inaccuracies in a chatbot, but the executive level would need to approve any deployment that affects customer data or regulatory compliance.

Policy implications extend beyond internal governance to external regulations. The European Union's AI Act classifies generative AI systems as general-purpose AI and imposes transparency obligations, including disclosure of training data sources and model capabilities. The framework maps these obligations to specific technical controls, such as the use of watermarks for generated content and the provision of model cards that summarize performance and limitations [21]. In the United States, the Office of Management and Budget's draft guidance on AI procurement similarly requires agencies to conduct risk assessments before acquiring AI systems. The framework can serve as a template for such assessments, providing a standardized yet flexible approach that can be adapted to different regulatory regimes.

Cross-border data flows and jurisdictional conflicts add another layer of complexity. Enterprises operating globally must reconcile differing requirements, such as the European Union's strict privacy rules under GDPR with the more permissive approaches in some Asian markets. The framework recommends a principle-based approach where the highest standard of risk mitigation is adopted across all jurisdictions to minimize legal exposure. Additionally, the framework encourages enterprises to participate in industry consortiums and standardization bodies to help shape emerging norms and best practices [22].

7. Deployment and Sustainability Considerations

The deployment of generative AI in enterprise information systems presents unique operational challenges. Unlike traditional software, which is typically updated through scheduled releases, generative models are often fine-tuned continuously based on user interactions, leading to potential drift in behavior. The framework mandates a rigorous model update policy that requires re-evaluation of risks whenever the model is modified, whether through fine-tuning, prompt changes, or alterations to the data pipeline. This policy includes a staging environment where updates are tested against a suite of risk metrics before being promoted to production.

Sustainability is increasingly recognized as an integral part of risk assessment. The training and inference of large generative models require substantial energy, which can contribute to carbon emissions and operational costs. Enterprises must consider the environmental impact when selecting model architectures and deployment strategies. For example, using smaller, distilled models for low-risk tasks can significantly reduce energy consumption without sacrificing acceptable performance. The framework encourages the adoption of green AI practices, such as using low-precision quantization and efficient transformer implementations

[23]. Furthermore, organizations should report their AI-related energy consumption as part of their corporate sustainability disclosures.

Model lifecycle management is another critical aspect. Generative models may become obsolete or unsafe as new vulnerabilities are discovered or as societal norms evolve. The framework requires that enterprises establish a retirement plan for each model, specifying triggers for decommissioning, data archiving, and migration to successor systems. This ensures that outdated models are not inadvertently left running, posing latent risks. Continuous monitoring systems should alert operators when model behavior deviates beyond predefined thresholds, enabling proactive intervention.

8. Conclusion

This paper has presented a comprehensive risk assessment framework for trustworthy generative AI in enterprise information systems, weaving together technical, organizational, and regulatory considerations. The framework addresses the unique challenges posed by generative models, including emergent behaviors, opaque reasoning, and dynamic risk profiles. By structuring the assessment around provenance, data lineage, and trustworthiness dimensions, it provides actionable guidance for enterprises seeking to deploy generative AI responsibly. The methodology's emphasis on iterative evaluation and continuous monitoring reflects the reality that risk is not static but evolves with model updates and changing contexts. The framework also highlights the importance of governance structures, policy alignment, and sustainability as integral components of risk management.

The proposed approach has limitations that suggest directions for future work. The framework is designed to be flexible, but its application requires significant organizational maturity in risk management and AI literacy. Small and medium-sized enterprises may find the comprehensive requirements daunting; thus, a simplified version or sector-specific adaptations could be beneficial. Additionally, the framework currently focuses on technical and procedural controls but does not fully address the societal dimensions of risk, such as the potential for generative AI to undermine public discourse or democratic processes. Future research could extend the framework to incorporate these broader impacts, drawing on insights from critical theory and political science. Furthermore, as generative AI technology continues to evolve, the framework must be periodically updated to reflect new capabilities and failure modes. Despite these limitations, the framework offers a systematic starting point for enterprises aiming to balance innovation with trustworthiness, providing a roadmap for responsible integration of generative AI into information systems.

References

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

4. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

5. Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 4691-4697.

6. Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

7. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. COM(2021) 206 final.

8. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

9. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 1050-1059.

10. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

11. Heaven, W. D. (2022). The new rules for AI. MIT Technology Review, 125(3), 24-31.

12. Hupont, I., & Charisi, V. (2022). AI risk management: A systematic literature review. IEEE Transactions on Technology and Society, 3(3), 167-182.

13. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

14. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

15. Klys, J., Snell, J., & Zemel, R. (2018). Learning latent subspaces for robust prediction and generation. Advances in Neural Information Processing Systems, 31.

16. Larson, J., Mattu, S., & Angwin, J. (2016). How we examined the COMPAS recidivism algorithm. ProPublica. https://www.propublica.org/article/how-we-examined-the-compas-recidivism-algorithm

17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

18. Microsoft Corporation. (2020). Responsible AI: A framework for building trusted AI systems. Microsoft AI Blog.

19. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. Proceedings of the 29th IEEE Symposium on Security and Privacy, 111-125.

20. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

21. Schneider, J., & Roos, A. (2021). Towards a certification framework for AI systems. AI and Ethics, 1(4), 395-407.

22. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.

23. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650.