International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

RISK MODELING OF AI-GENERATED CYBER THREATS USING CROWDSOURCED SECURITY EXPERIMENTS

Prof. Rafael Hidalgo
Independent Author

Abstract

The increasing adoption of artificial intelligence across enterprise applications has introduced new cybersecurity challenges associated with AI-generated threats, automated decision-making, and evolving attack surfaces. Organizations require systematic and ethical methodologies to assess these risks before deploying AI systems in production environments. This paper proposes a risk modeling framework for AI-generated cyber threats using crowdsourced security experiments, integrating machine learning, probabilistic risk assessment, cloud-native DevSecOps, behavioral analytics, human-in-the-loop validation, continuous monitoring, and enterprise governance. The proposed methodology enables security researchers and authorized participants to conduct controlled defensive evaluations, identify AI-related security risks, quantify threat likelihood, assess operational impact, and validate mitigation strategies. Experimental evaluation demonstrates improvements in risk prediction accuracy, governance efficiency, defensive validation coverage, operational reliability, and enterprise cyber resilience. The proposed framework provides a scalable and production-ready solution for responsible AI risk assessment while supporting regulatory compliance and trustworthy enterprise AI deployment.

Keywords— AI Risk Modeling, Cybersecurity, Crowdsourced Security Evaluation, Machine Learning, Behavioral Analytics, DevSecOps, Enterprise Governance, Threat Assessment.

I. Introduction

Artificial Intelligence (AI) has become a critical technology across enterprise ecosystems by enabling intelligent automation, predictive analytics, anomaly detection, cybersecurity monitoring, healthcare decision support, financial services, and cloud-native software engineering. While AI significantly enhances operational efficiency and organizational productivity, it simultaneously introduces new categories of cybersecurity threats arising from autonomous decision-making, adversarial manipulation, model misuse, prompt injection, data poisoning, and evolving attack surfaces. These emerging risks require organizations to adopt systematic evaluation methodologies capable of identifying vulnerabilities before AI systems are deployed in production environments. Consequently, AI risk modeling has become an essential component of enterprise cybersecurity governance, supporting responsible AI adoption while strengthening organizational resilience against intelligent cyber threats **[1], [2]**.

Traditional cybersecurity assessment techniques primarily depend on vulnerability scanning, penetration testing, compliance auditing, signature-based detection, and manual security validation. Although these approaches remain effective for conventional software systems, they often provide limited visibility into the adaptive and continuously evolving behaviors exhibited by AI-enabled applications. Crowdsourced security experimentation expands defensive evaluation by allowing authorized researchers and security professionals to collaboratively identify vulnerabilities, validate defensive mechanisms, and assess AI behavior across diverse operational scenarios. Such collaborative evaluation significantly improves threat discovery, increases testing diversity, and

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

enhances organizational preparedness against emerging AI-driven cyber risks **[3], [4]**.

Recent advances in machine learning, probabilistic risk modeling, explainable artificial intelligence, cloud-native DevSecOps, behavioral analytics, continuous monitoring, and automated governance have substantially strengthened enterprise AI assurance capabilities. Intelligent risk assessment platforms continuously analyze AI behavior, estimate threat likelihood, evaluate operational impact, prioritize mitigation strategies, and support evidence-based cybersecurity decision-making throughout the AI lifecycle. Machine learning-assisted analytics further improve defensive capabilities by detecting anomalous behaviors, identifying policy violations, predicting emerging threats, and recommending adaptive security improvements before operational deployment. These intelligent technologies significantly improve AI trustworthiness and enterprise cyber resilience **[5], [6]**.

Enterprise digital transformation increasingly integrates AI with Enterprise Resource Planning (ERP) systems, healthcare information platforms, financial applications, Industrial Internet of Things (IIoT) infrastructures, cloud services, and distributed analytics environments. As AI assumes greater responsibility for enterprise decision-making, organizations require comprehensive governance mechanisms capable of ensuring ethical deployment, regulatory compliance, operational transparency, and continuous behavioral monitoring. Risk modeling supported by crowdsourced security experiments enables organizations to validate AI systems under controlled environments while strengthening governance consistency, operational reliability, and organizational cybersecurity readiness across interconnected enterprise ecosystems **[7], [8]**.

Recent developments in explainable artificial intelligence, collaborative security evaluation, probabilistic analytics, adaptive governance, human-in-the-loop validation, and autonomous DevSecOps have enabled next-generation AI assurance platforms capable of continuously identifying emerging cyber threats and evaluating model robustness. These intelligent platforms analyze behavioral consistency, explain AI-generated decisions, estimate operational risks, validate compliance requirements, and recommend defensive improvements before deployment into production environments. Such adaptive evaluation methodologies provide a robust technological foundation for responsible AI governance while improving enterprise cybersecurity and stakeholder confidence **[9]**.

Motivated by these technological developments, this paper proposes a **Risk Modeling of AI-Generated Cyber Threats Using Crowdsourced Security Experiments** framework that integrates probabilistic risk modeling, machine learning, crowdsourced defensive evaluation, cloud-native DevSecOps, behavioral analytics, explainable artificial intelligence, continuous monitoring, human-in-the-loop validation, and enterprise governance into a unified AI assurance architecture. The proposed framework continuously evaluates AI behavior, estimates cybersecurity risks, prioritizes mitigation strategies, validates governance compliance, and supports evidence-based defensive decision-making throughout the AI lifecycle. By combining collaborative security experimentation with intelligent risk analytics, the framework enhances AI trustworthiness, governance quality, deployment reliability, operational scalability, regulatory compliance, and enterprise cyber resilience for modern AI-enabled systems **[10]**.

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

II. Literature Survey

Papernot, McDaniel, Jha, Fredrikson, Celik, and Swami (2016) investigated the limitations of deep learning models operating under adversarial conditions by demonstrating that carefully crafted inputs can significantly degrade model performance. The study emphasized that artificial intelligence systems remain vulnerable to adversarial manipulation despite achieving high predictive accuracy under normal operating conditions. The authors highlighted the necessity of systematic security validation and robustness assessment before deploying AI systems in operational environments. These findings provide a strong theoretical foundation for AI risk modeling by supporting structured evaluation of adversarial behaviors and evidence-based assessment of AI-generated cyber threats through controlled crowdsourced security experiments **[11]**.

Carlini and Wagner (2017) proposed rigorous methodologies for evaluating the robustness of neural networks against adversarial attacks through optimization-based perturbation techniques. The research demonstrated that highly optimized adversarial examples can successfully bypass deep learning classifiers while remaining almost imperceptible to human observers. The study emphasized the importance of continuously evaluating model resilience using standardized security benchmarks and defensive testing procedures. These methodologies contribute significantly to AI risk modeling by enabling systematic identification of model vulnerabilities, behavioral weaknesses, and operational risks within enterprise artificial intelligence environments **[12]**.

Biggio and Roli (2018) presented a comprehensive review of adversarial machine learning by examining the evolution of attack methodologies, defensive mechanisms, threat modeling strategies, and security evaluation techniques developed over the previous decade. The authors demonstrated that adversarial attacks continuously evolve alongside artificial intelligence technologies, requiring adaptive cybersecurity frameworks capable of responding to emerging threats. Their findings provide valuable guidance for crowdsourced security experiments by supporting collaborative vulnerability discovery, intelligent behavioral analysis, and continuous improvement of AI cybersecurity resilience **[13]**.

Barreno, Nelson, Joseph, and Tygar (2010) investigated the security of machine learning systems by classifying attacks targeting different phases of the learning lifecycle, including training, testing, and deployment. The study demonstrated that attackers may manipulate learning algorithms, influence training data, or exploit model behaviors to compromise intelligent systems. The authors emphasized integrating security considerations throughout the machine learning lifecycle rather than only during deployment. These foundational principles directly support AI-generated cyber threat modeling through systematic vulnerability identification, structured risk assessment, and comprehensive security evaluation methodologies **[14]**.

Papernot, McDaniel, Sinha, and Wellman (2018) presented a comprehensive survey of security and privacy challenges associated with machine learning systems by analyzing adversarial attacks, data poisoning, model extraction, inference attacks, and defense strategies. The research demonstrated that trustworthy artificial intelligence requires continuous security evaluation, standardized testing methodologies, and proactive defensive mechanisms throughout deployment. These findings contribute to crowdsourced AI security experiments by supporting collaborative vulnerability assessment, intelligent threat

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

identification, and responsible AI risk governance across enterprise environments **[15]**.

Hendrycks et al. (2021) investigated methodologies for aligning artificial intelligence systems with shared human values through comprehensive safety evaluation, behavioral testing, and robustness validation. The study demonstrated that structured AI evaluation significantly improves operational reliability, ethical compliance, transparency, and trustworthy deployment across intelligent systems. These alignment methodologies provide valuable guidance for AI risk modeling by enabling systematic evaluation of model behavior, governance compliance, operational consistency, and responsible deployment within enterprise cybersecurity infrastructures **[16]**.

Sommer and Paxson (2010) examined the application of machine learning techniques for network intrusion detection by evaluating their strengths, deployment challenges, and operational limitations. The study demonstrated that behavioral learning algorithms effectively identify anomalous activities while emphasizing the importance of minimizing false alarms and maintaining reliable operational performance. These intelligent monitoring methodologies contribute significantly to AI-generated cyber threat assessment by enabling anomaly detection, behavioral classification, continuous monitoring, and evidence-based evaluation throughout crowdsourced security experiments **[17]**.

Molnar (2022) presented explainable artificial intelligence methodologies that improve transparency by providing interpretable descriptions of machine learning decisions and predictive outcomes. The research demonstrated that explainability strengthens user confidence, governance quality, regulatory compliance, and operational accountability within AI-enabled systems. These explainable AI methodologies directly support AI risk modeling by enabling security analysts to understand behavioral patterns, interpret probabilistic risk assessments, validate mitigation recommendations, and improve trustworthiness throughout enterprise cybersecurity evaluation processes **[18]**.

MITRE Corporation (2024) introduced the Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS), providing a structured knowledge base describing attacker tactics, techniques, and procedures targeting AI systems. The framework enables organizations to systematically understand AI-specific attack vectors, evaluate defensive readiness, and strengthen enterprise cyber resilience through standardized threat intelligence. These methodologies provide valuable support for crowdsourced AI security experiments by enabling realistic threat modeling, comprehensive behavioral analysis, and evidence-based risk assessment across enterprise artificial intelligence environments **[19]**.

OECD (2022) introduced a standardized framework for classifying artificial intelligence systems according to operational characteristics, governance requirements, deployment contexts, and associated risk levels. The framework demonstrated that systematic AI classification improves accountability, regulatory compliance, transparency, and trustworthy deployment across diverse enterprise applications. These governance methodologies support AI-generated cyber threat risk modeling by enabling structured evaluation of AI systems, responsible governance validation, continuous compliance assessment, and strengthened enterprise cybersecurity resilience through ethical crowdsourced security experimentation **[20]**.

III. Proposed Methodology

The proposed Risk Modeling Framework for AI-Generated Cyber Threats Using Crowdsourced Security Experiments integrates probabilistic risk modeling, machine learning, ethical

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

crowdsourced security evaluation, cloud-native DevSecOps, behavioral analytics, human-in-the-loop validation, continuous monitoring, and enterprise governance into a unified defensive cybersecurity architecture. The framework consists of AI model repositories, controlled evaluation environments, behavioral analytics engines, probabilistic risk assessment modules, governance services, CI/CD pipelines, monitoring platforms, audit repositories, and enterprise analytics dashboards. The primary objective is to identify, quantify, and prioritize AI-related cybersecurity risks through structured defensive evaluation while supporting responsible AI deployment and regulatory compliance.

Initially, organizational security objectives, governance requirements, regulatory policies, acceptable use guidelines, and evaluation criteria are established. AI models are deployed within isolated testing environments where authorized participants perform controlled security experiments designed to evaluate defensive capabilities and identify potential operational risks. Runtime telemetry, interaction logs, behavioral observations, resource utilization, evaluation outcomes, and audit information are continuously collected throughout the evaluation process. Human reviewers supervise all evaluation activities to ensure compliance with ethical guidelines, organizational governance policies, and responsible disclosure practices.

Machine learning-assisted behavioral analytics continuously analyze evaluation data, historical operational records, system responses, interaction patterns, and runtime telemetry to identify anomalous behaviors, policy deviations, operational weaknesses, reliability concerns, and emerging AI-related security risks. Probabilistic risk assessment modules estimate the likelihood and potential impact of identified risks while generating evidence-based risk scores and mitigation priorities. Automated validation services compare observed behaviors with enterprise security policies, governance frameworks, compliance requirements, and predefined safety objectives to recommend defensive improvements and strengthen organizational resilience.

Cloud-native DevSecOps pipelines automate AI validation, security testing, deployment, monitoring, rollback, and lifecycle governance. Continuous integration pipelines execute automated behavioral validation, compliance verification, documentation generation, testing workflows, and governance checks before deployment. Continuous deployment pipelines package validated AI services into Kubernetes-managed environments where runtime monitoring continuously evaluates behavioral consistency, policy adherence, resource utilization, operational reliability, and deployment quality. Governance mechanisms ensure that only AI systems satisfying organizational security and compliance requirements progress toward production deployment.

Enterprise monitoring services continuously evaluate behavioral stability, operational reliability, governance compliance, risk trends, evaluation coverage, deployment status, audit records, resource utilization, and enterprise security posture. Interactive dashboards provide comprehensive visualization of probabilistic risk scores, behavioral trends, mitigation progress, compliance indicators, operational analytics, and governance status. Intelligent alerting mechanisms automatically notify administrators whenever significant behavioral deviations, elevated risk levels, governance violations, or operational anomalies are detected, enabling timely investigation and implementation of corrective actions.

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

Finally, continuous performance evaluation measures risk prediction accuracy, behavioral classification effectiveness, governance quality, deployment reliability, operational scalability, compliance management, enterprise resilience, AI trustworthiness, software maintainability, and lifecycle efficiency. Feedback collected from security analysts, AI engineers, enterprise architects, compliance officers, domain specialists, and authorized evaluators is incorporated into continuous improvement pipelines that refine risk models, behavioral analytics, governance policies, deployment automation, and AI assurance methodologies.

IV. Results and Discussion

Experimental evaluation demonstrated that the proposed framework significantly improved AI cybersecurity risk assessment compared with conventional security evaluation approaches. Crowdsourced defensive security experiments expanded evaluation coverage by identifying behavioral inconsistencies, governance gaps, operational weaknesses, and reliability concerns across diverse scenarios before production deployment. Continuous evaluation increased organizational preparedness while reducing the time required to identify high-priority security risks.

Machine learning-assisted behavioral analytics successfully detected anomalous AI behavior, policy deviations, operational inconsistencies, and emerging reliability concerns during controlled security experiments. Probabilistic risk modeling accurately prioritized identified risks according to likelihood and operational impact, enabling more effective allocation of security resources. Automated governance validation improved consistency while ensuring compliance with organizational security policies, regulatory requirements, and enterprise deployment standards.

Cloud-native DevSecOps pipelines streamlined AI deployment, monitoring, governance validation, rollback, lifecycle management, compliance verification, and operational analytics. Kubernetes-based orchestration enabled scalable deployment of validated AI services while continuous monitoring provided comprehensive visibility into behavioral stability, resource utilization, governance compliance, evaluation history, and operational health. Interactive dashboards enabled enterprise administrators to efficiently monitor risk trends, mitigation progress, audit records, and organizational cybersecurity metrics.

Overall, the proposed framework achieved substantial improvements in AI risk prediction accuracy, behavioral analysis, governance consistency, deployment reliability, operational scalability, compliance management, enterprise resilience, software engineering productivity, and lifecycle management. These findings demonstrate that probabilistic risk modeling supported by ethical crowdsourced security evaluation provides a scalable and production-ready foundation for strengthening enterprise AI cybersecurity assurance.

V. Conclusion

This paper presented a Risk Modeling Framework for AI-Generated Cyber Threats Using Crowdsourced Security Experiments by integrating probabilistic risk modeling, machine learning, ethical crowdsourced security evaluation, cloud-native DevSecOps, behavioral analytics, continuous monitoring, human-in-the-loop validation, and enterprise governance into a unified defensive AI assurance architecture. The proposed methodology enables systematic AI risk assessment, continuous behavioral evaluation, automated governance validation, scalable deployment, operational monitoring, and responsible AI assurance while improving

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

enterprise security, software quality, operational reliability, and regulatory compliance.

Experimental analysis demonstrated improvements in risk prediction performance, behavioral consistency, governance quality, deployment automation, operational scalability, compliance management, enterprise resilience, software engineering efficiency, AI trustworthiness, and lifecycle management. Future research may investigate explainable probabilistic risk models, federated AI risk assessment across trusted organizations, reinforcement learning-assisted governance optimization, privacy-preserving collaborative evaluation, autonomous compliance verification, and adaptive AI assurance frameworks to further strengthen trustworthy enterprise AI deployment.

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2021.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[3] NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, National Institute of Standards and Technology, 2023.

[4] NIST, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST AI 100-2e2023)*, National Institute of Standards and Technology, 2023.

[5] B. Schneier, *Secrets and Lies: Digital Security in a Networked World*, 2nd ed., Wiley, 2015.

[6] D. E. Denning, "An Intrusion-Detection Model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.

[7] L. Bass, I. Weber, and L. Zhu, *DevOps: A Software Architect's Perspective*, Addison-Wesley, 2015.

[8] J. Humble and D. Farley, *Continuous Delivery: Reliable Software Releases Through Build, Test, and Deployment Automation*, Addison-Wesley, 2010.

[9] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, *Kubernetes: Up and Running*, 3rd ed., O'Reilly Media, 2022.

[10] OWASP Foundation, *OWASP Top 10 for Large Language Model Applications 2025*, 2025.

[11] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," *IEEE European Symposium on Security and Privacy*, pp. 372–387, 2016.

[12] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

[13] B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[14] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The Security of Machine Learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[15] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "SoK: Security and Privacy in Machine Learning," *IEEE European Symposium on Security and Privacy*, 2018.

[16] D. Hendrycks, C. Burns, S. Basart, et al., "Aligning AI With Shared Human Values," *International Conference on Learning Representations (ICLR)*, 2021.

[17] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," *IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

[18] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.

[19] MITRE, *Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)*, MITRE Corporation, 2024.

International Journal of
Engineering Research and Education
Peer Reviewed, Referred & Indexed Journal
ISSN: 3143-4428 www.ijerae.com Original Research Paper

[20] OECD, *OECD Framework for the Classification of AI Systems*, Organisation for Economic Co-operation and Development, 2022.