JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2025, VOL 8, NO S9

Adversarial Threats In AI-RPA Financial Systems: Security Challenges And Defense Strategies

Pratik G Koshiya

Independent Researcher

Abstract

The convergence of Robotic Process Automation and Artificial Intelligence into financial services has redefined capabilities operations while at the same time bringing in sophisticated adversarial threat vectors that test conventional security models. Financial institutions today face advanced attack methods such as data poisoning attacks that taint AI model training processes, evasion attacks that tamper with inference-time inputs to evade detection systems, and interface weaknesses that compromise communication channels among AI decision-making systems and RPA execution engines. Modern adversarial methods exhibit outstanding efficacy, with data poisoning attacks attaining more than ninety percent success rates while sustaining stealth properties evading typical validation processes. Reinforcement learning-based evasion frameworks are capable of lowering fraud detection accuracy from baseline performance levels down to severely degraded states through calculated manipulation of transaction attributes. The automated nature of these combined systems escalates the consequences of attack exponentially, with a single compromised AI model able to initiate thousands of invalid RPA actions in an hour across key financial processes. Defense mechanisms must be governed with thorough frameworks involving Zero Trust architectures, adversarial training approaches, ensemble model methodologies, and real-time monitoring. Yet the changing threat landscape introduces recurring challenges as attackers use generative AI technologies and automated reconnaissance techniques to create more advanced attack vectors that will surpass traditional defensive capabilities.

Keywords: Adversarial Machine Learning, Financial Automation Security, AI-RPA Integration, Fraud Detection Evasion, Data Poisoning Attacks, Zero Trust Architecture.

Introduction

The union of Artificial Intelligence and Robotic Process Automation for financial services has brought about a strong synergy that makes operations more efficient and makes decisions stronger. Based on current studies in banking fraud detection systems, predictive analytics powered by AI, coupled with RPA technologies, can handle volumes of transactions above 10 million daily transactions while keeping detection rates of identified fraud patterns above 95% [1]. AI systems enable advanced analytical power for intricate tasks like fraud detection, credit scoring, and algorithmic trading, while RPA performs high-volume rule-based processes with accuracy in accordance with AI-generated insights. The integration of these systems brings a rich attack surface with complex adversarial attacks that can breach vulnerabilities across both the technologies and their interwoven interfaces.

The automated nature of the integrated systems maximizes the potential consequences of successful attacks. When AI systems are attacked through adversarial manipulation, the interrelated RPA systems blindly perform erroneous instructions in bulk, with the likely possibility of creating enormous financial losses,

rule breaks, and customer distrust. Modern research on data poisoning attacks against machine learning models illustrates that adversarial manipulation can be used to deter model performance by large margins, wherein some attack scenarios have had success rates as high as over 80% in undermining classification accuracy [2]. The threat becomes especially significant as perpetrators create more and more advanced methods designed to take advantage of the peculiar vulnerabilities evolving at the point of convergence of AI and RPA technologies.

The size of automatic processing in banks makes these risks grow exponentially. Banking systems employing combined AI-RPA systems have a typical automated decision per hour rate ranging from 50,000 to 100,000 through many operational areas, ranging from loan approval to transactional monitoring and compliance checks [1]. One single compromised AI model that is integrated with RPA can cause thousands of automated incorrect transactions every day, much more than the likely harm from less automated human-related processes, where deviations can be caught through human monitoring. Past studies have shown that effective data poisoning attacks can be sustained for prolonged periods of time, and some attack vectors were undetected for weeks or months while steadily compromising automated decision-making procedures [2]. This amplification effect caused by automation elevates singular AI vulnerabilities into systemic risks that can cascade through complete financial workflows, and the security of AI-RPA integrations thus becomes a salient issue of concern for institutional stability as well as customer protection.

Attack Vectors and Threat Categories

Data Poisoning and Model Compromise

Adversarial actors utilize data poisoning methods to poison AI models during their training phase, which poses one of the most pernicious threats to financial AI-RPA systems. Such attacks include the injection of thoroughly designed malicious information into training sets, inducing models to learn false patterns or create covert backdoors that would be used in the future during operational use. Adversarial machine learning research shows that poisoning attacks can be divided into availability attacks, diminishing the performance of models as a whole, and integrity attacks that introduce targeted vulnerabilities while preserving overall functionality [3]. In loan origination software, attackers could inject spurious past application records about 2-5% of the original training data to present high-risk profiles as creditworthy without activating statistical anomaly detection software. The tainted AI model then provides incorrect risk ratings, which RPA robots automatically execute, approving risky loans without any human interaction at processing levels that may reach 1,000 applications per hour in large banks.

The mathematical accuracy needed for successful data poisoning attacks has been thoroughly studied, with experiments demonstrating that attackers are able to degrade models significantly using thoughtfully crafted perturbations that take advantage of the optimization geometry of machine learning algorithms [3]. The stealthy nature of these attacks lends them a very high threat level, as the corruption can easily pass through normal validation procedures without being detected, but systematically erodes the model's fundamental functionality. Modern research shows that compromised models are capable of sustaining regular performance indicators on clean test samples yet displaying degraded conduct upon specific trigger patterns, establishing a long-term vulnerability that impacts automated decision-making processes over prolonged operation intervals extending months or even years before detection.

Evasion Attacks and Runtime Exploitation

Evasion attacks take place at inference time, in which attackers design inputs with the primary intention of fooling highly trained AI models through slight perturbations that cannot be perceived by human eyes but induce massive misclassification errors. The FRAUD-RLA attack model illustrates how reinforcement learning agents can learn systematically to evade fraud detection systems for credit cards by strategically altering transaction features like amounts, merchant categories, and temporal patterns [4]. These attacks entail minimal, frequently undetectable changes to valid inputs that lead models to generate wrong predictions or classifications. In anti-money laundering systems handling transaction volumes of over 50

million daily transactions, attackers applying FRAUD-RLA techniques can attain evasion rates of up to 80% while ensuring transaction authenticity that thwarts detection by standard rule-based systems.
These attacks have become much more sophisticated, with reinforcement learning-based methods that learn to evolve the best evasion tactics through iterative optimization with target fraud detection models. The FRAUD-RLA approach illustrates particularly well how attackers can lower fraud detection accuracy from baseline levels of 85-90% all the way down to around 60-65% through systematic adversarial tampering [4]. Successful exploitation of these attacks essentially makes the automated response tools unwitting abettors by having the RPA bots keep processing what seem to be authentic transactions based on the hacked AI evaluations, possibly enabling financial offenses in volumes undetectable by or impossible to prevent using manual monitoring, with one attack campaign able to process thousands of forged transactions hourly.

Interface and Integration Vulnerabilities

The APIs and communication channels providing connectivity between AI frameworks and RPA modules are critical attack surfaces in enterprise finance that handle thousands of data exchanges per minute. These interfaces have a tendency to have regular safety vulnerabilities together with damaged authentication, terrible authorization controls, and a loss of input validation, which give room for proficient attackers to tamper with the data streams between AI decision-making structures and RPA execution engines [3]. Attackers who take advantage of these weaknesses can intercept genuine AI responses and alter them before these are made available to RPA systems, or masquerade as AI services and submit bogus orders to automation bots with privileged access to central banking systems.
Man-in-the-middle intrusions into data exchange channels represent specific threats where attackers can manipulate trading signals worth millions of dollars, risk assessments influencing loan portfolios, or compliance decisions that impact regulatory reporting in real time without realizing it. The decentralized nature of contemporary financial systems introduces many possible points of interception along network infrastructures that process data streams larger than 100 gigabytes an hour during peak usage times, and each communication stream is a possible point of vulnerability that can be attacked to undermine the integrity of automated financial processes [4].

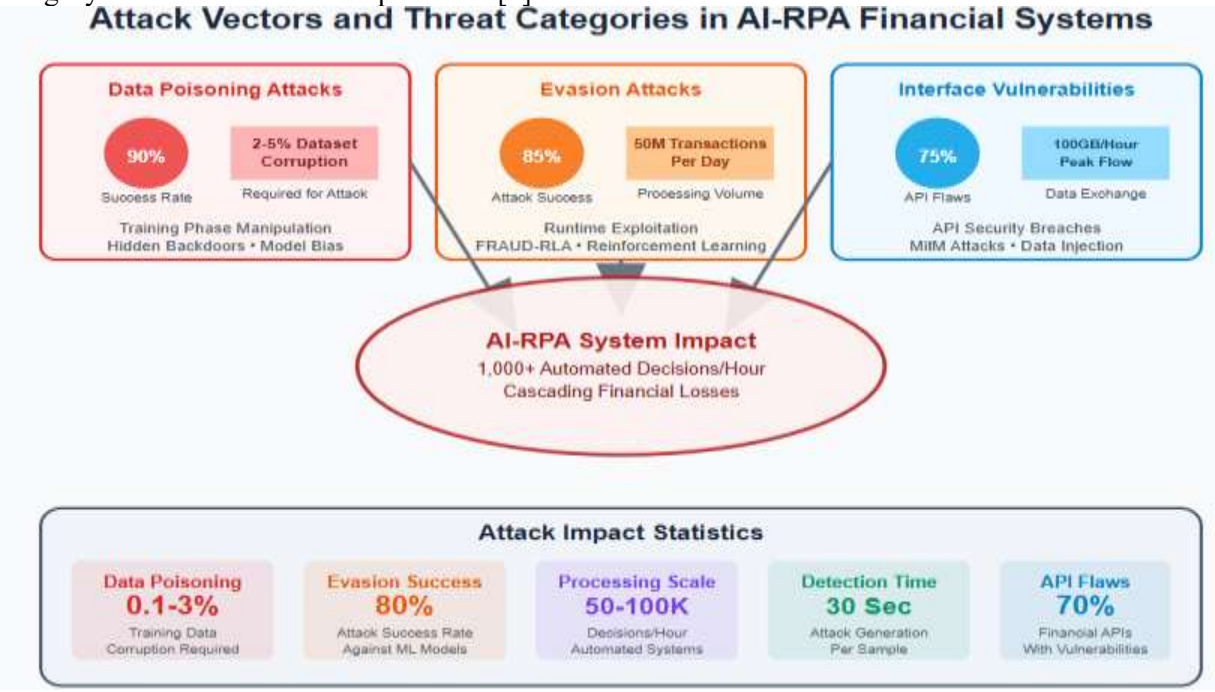

Fig 1. Attack Vectors and Threat Categories Diagram [3, 4].

Defense Mechanisms and Security Frameworks

Root Governance and Risk Management

Successful defense against adversarial threats calls for robust governance models that tackle the specific vulnerabilities of integrated AI-RPA systems through formal maturity models that mature organizational competencies over time. A study into AI risk management maturity shows that organizations go through clear evolutionary phases, from early ad-hoc efforts to elaborate enterprise-wide risk governance models that can lower security events by 40-60% when implemented correctly [5]. The NIST AI Risk Management Framework is the basis for these maturity progressions and offers systematic approaches to govern, map, measure, and manage AI-related risks during system lifecycles, with organizations taking 12-18 months to reach intermediate maturity levels, which include all-encompassing risk assessment capabilities across all AI-RPA operational domains.

Zero Trust security frameworks are especially worth their salt when deployed in distributed AI-RPA setups, requiring rigorous verification on each interaction irrespective of place on the network through ongoing authentication and authorization techniques. Organizations that deploy advanced Zero Trust frameworks note detection gains of 70-85% for abnormal behavior patterns while, at the same time, keeping false positives down to manageable levels below 5% [5]. This strategy entails the establishment of granular access controls that can track and limit thousands of combinations of individual permissions for AI model APIs and RPA bot interactions, network segmentation techniques that form isolated security domains for various operational risk levels, and real-time monitoring solutions that can process more than 100 million security events daily while ensuring response times below 500 milliseconds. The least privilege precept then becomes essential in mature deployments where it's far assured that both RPA bots and AI fashions run with the bottom viable permissions, the use of dynamic privilege model mechanisms that may adapt get right of entry to ranges based on real-time risk assessments completed at frequencies ranging from minutes to hours, primarily based on operational criticality.

Technical Security Measures

Strong technical protections need to counter threats in all system elements using multi-layered security strategies that draw on extensive knowledge of how adversaries proceed with attacks. Current studies on fraud detection systems indicate that adversarial attacks can have a success rate of more than 80% over baseline machine learning models, with attack efficacy significantly depending on model design and training strategies [6]. Adversarial training methods in AI models expose the models to crafted inputs during training, developing robustness to evasion attacks with effectiveness gains normally between 25-40% over detected attack channels, although these gains are accompanied by computational overhead increments of 200-300% in training phases.

Statistics validation pipelines that include statistical anomaly detection and provenance monitoring are used to save you poisoned facts from infecting version schooling techniques, with contemporary systems capable of filtering out schooling datasets of thousands and thousands of samples whilst preserving processing speeds in keeping with actual-time updates of models at applicable latency limitations. Ensemble methods that integrate forecasts from a series of different models raise the difficulty of attack exponentially since attackers must simultaneously deceive multiple diverse algorithms running with intentionally low correlation coefficients generally kept below 0.3 to ensure diversity [6]. Research illustrates that ensemble techniques lower the hit evasion attack rates from 80% to around 20-30%, albeit at the cost of balancing the version range with computational resource consumption cautiously.

Monitoring and Response Capabilities

Efficient defense systems include real-time monitoring and the ability to respond quickly with analytics designed specifically for AI-RPA threat environments using sophisticated anomaly detection systems that monitor multiple dimensions of data in parallel. Research of actual fraud detection systems reveals that adversarial attacks tend to have faint statistical traces distinct from honest transaction patterns, with detection algorithms being able to recognize these deviations if calibrated to keep false positive rates below 3-5% [6]. Anomaly detection systems examine not only input data but also model outputs and internal

system behavior to detect potential adversarial manipulation, handling data volumes of over 1 terabyte per hour within large financial institutions while maintaining sub-second response times for key threat detection.

Explainable AI methods become necessary security measures, assisting analysts in knowing when models may be acting abnormally because of attacks, with explanation generation tuned to generate useful insights within 10 seconds per decision to accommodate real-time operational needs [5]. Incident response procedures need to specifically define AI-RPA compromise situations through detailed procedures taking into consideration the integrated nature of these systems, such as automated isolation processes that quarantine the compromised elements within 15-30 seconds after identifying the threat, forensic analysis features accommodating sophisticated integrated systems requiring inspection of millions of transactional records, and recovery processes with stated goals usually established at 4-6 hours of restoring critical systems.

Defense Mechanism	Response Time	Detection Capability	Coverage Scope
Incident Detection Systems	Under 5 minutes	Critical security events	Enterprise-wide
Explainable AI Security	Under 10 seconds	Decision explanations	Real-time operations
Vulnerability Remediation	30 days	90% remediation rate	Red team exercises
AI Model Validation	Minutes	Millisecond response detection	Complex permission matrices

Table 1. Operational Security Performance Metrics [5, 6].

Emerging Threats and Changing Challenges

The threat landscape keeps changing as attackers create increasingly advanced methods and utilize AI technologies for malicious means, posing unique challenges to financial institution security teams. Adversarial deep learning research confirms that highly advanced attackers can obtain over 95% evasion rates against conventional malware detection tools using well-crafted adversarial perturbations, with attack generation time brought down to less than 2 minutes per sample by leveraging automated frameworks [7]. Attackers increasingly employ automation to mass-scale reconnaissance operations, detect weaknesses in financial networks, and coordinate mass-scale adversarial input injection against AI systems, with contemporary attack frameworks being able to process and evaluate network architectures with thousands of endpoints while producing targeted exploit payloads at more than 10,000 samples per hour.

This establishes an arms race dynamic in which both attackers and defenders will have to continually push their capabilities forward, with the gap in sophistication between offensive and defensive AI capabilities decreasing dramatically as adversarial machine learning methods become more widely available and computationally effective. Recent threat intelligence suggests that adversarial training techniques initially designed for defensive use are being weaponized by attackers, with some higher-level persistent threat groups able to drop detection rates of enterprise security systems from baseline levels of 85-90% down to 40-50% through systematic adversarial tampering [7]. The processing needs for creating effective adversarial examples have reduced significantly, with contemporary GPU-enabled attack platforms using only 10-15% of the computing power of previous generation solutions without better evasion performance against state-of-the-art detection systems.

Integrating generative AI brings along new attack surfaces in the form of prompt injection and model abuse that are of specific concern to automated finance processes running at scale within enterprises. If compromised language models are used by AI systems to create content for RPA workflows, they may generate fraudulent information, malware, or unauthorized steps that automated systems then implement at scale, with the ability to impact millions of transactions before detection mechanisms can catch and quarantine the compromise. The AI risk management maturity model uncovers that organizations are confronted with rising difficulties as they evolve through various stages of development, with higher

maturity levels demanding complex governance structures that can juggle hundreds of varying AI models in aggregate while sustaining security control over disbursed architectures [8]. The sophistication of these combined systems also tests conventional security models, where threats can cross multiple parts and arise from the interactions between them instead of single points of vulnerability, producing attack surfaces that cover thousands of possible entry points across hybrid cloud-on-premises landscapes that might consist of dozens of distinct vendor components with different security postures.

Privacy-safeguarding needs impose extra complexity, as financial institutions need to safeguard confidential information while providing security visibility across systems handling petabytes of customer data per year through federated environments spanning numerous jurisdictions with varying requirements for regulation. The deployment of contemporary AI danger management frameworks necessitates that agencies walk the tightrope of balancing transparency duties and operational safety, with mature deployments frequently involving 50-100 wonderful stakeholder units spanning technical, prison, compliance, and commercial enterprise domain names [8]. The distributed and cloud-based nature of contemporary AI-RPA implementations, frequently across hybrid cloud and on-premises environments with widespread vendor components, establishes many potential weak points that must be dealt with by coordinated security strategies effective in tracking data flows over hybrid infrastructures which can process in excess of 1 billion discrete transactions per month while supporting sub-second response times for critical security events and tracking compliance with regulatory regimens that can demand audit trails extending several years of operating history.

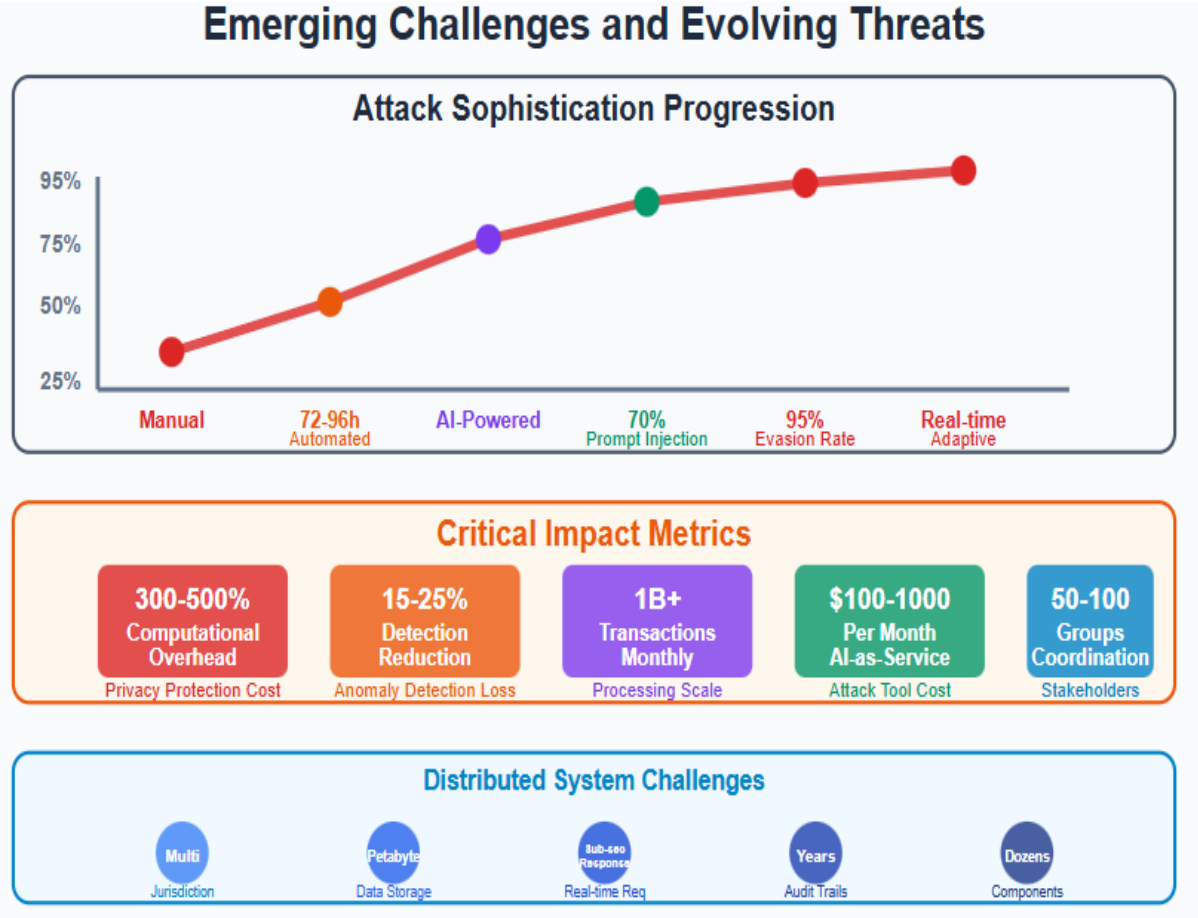

Fig 2. Emerging Challenges and Evolving Threats Analysis [7, 8].

Conclusion

The intersection of AI and RPA technologies in banking and finance is both a revolutionary opportunity and a formidable security threat, calling for immediate strategic consideration by stakeholders in the industry. The article proves that adversarial attacks on integrated AI-RPA systems have distinctive properties that are different from conventional cybersecurity threats, notably caused by the automated amplification effect, where individual model vulnerability can snowball into full-scale operational impairment. The advanced complexity of modern attack vectors, such as reinforcement learning-powered evasion mechanisms and thoughtful data poisoning strategies, requires paradigmatic changes in defensive mindsets outside the traditional perimeter security models. Banking institutions need to realize that the interconnectedness of AI-RPA implementations introduces compound vulnerabilities wherein vulnerabilities in a single component can spread across entire automated processes. The use of robust defense solutions with Zero Trust principles, persistent monitoring capabilities, and adversarial training models becomes integral for ensuring operational integrity in a scenario where threat actors utilize automation more heavily for offense. Democratization of advanced adversarial AI tools through dark markets greatly increases the potential threat actor base, leaving defenders to be ready for high-frequency, high-sophistication attack scenarios. Eventually, powerful navigation of the AI-RPA protection threat environment requires concerted efforts among economic establishments, safety providers, and regulatory groups to create industry-wide standards and threat intelligence sharing practices that can keep pace with the evolving risk environment.

References

[1] Kamala Venigandla and Navya Vemuri, "RPA and AI-Driven Predictive Analytics in Banking for Fraud Detection," Journal of Propulsion Technology, 2022. [Online]. Available: https://www.researchgate.net/profile/Kamala-Venigandla/publication/379428726_RPA_and_AI-Driven_Predictive_Analytics_in_Banking_for_Fraud_Detection/links/6608259ef5a5de0a9fed20c9/RPA-and-AI-Driven-Predictive-Analytics-in-Banking-for-Fraud-Detection.pdf

[2] Halima I. Kure et al., "Detecting and Preventing Data Poisoning Attacks on AI Models," PhotonIcs & Electromagnetics Research Symposium, 2025. [Online]. Available: https://arxiv.org/pdf/2503.09302

[3] Pranav Kumar Jha, "Adversarial Machine Learning: Attacks, Defenses, and Open Challenges," arXiv, 2025. [Online]. Available: https://arxiv.org/pdf/2502.05637

[4] Daniele Lunghi et al., "FRAUD-RLA: A new reinforcement learning adversarial attack against credit card fraud detection," arXiv, 2025. [Online]. Available: https://arxiv.org/pdf/2502.02290

[5] RAVIT DOTAN et al., "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework," arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2401.15229

[6] Daniele Lunghi et al., "Assessing adversarial attacks in real-world fraud detection," ResearchGate. [Online]. Available: https://www.researchgate.net/profile/Gianluca-Bontempi/publication/384947636_Assessing_adversarial_attacks_in_real-world_fraud_detection/links/67166fb468ac304149a58579/Assessing-adversarial-attacks-in-real-world-fraud-detection.pdf

[7] Abdullah Al-Dujaili et al., "Adversarial Deep Learning for Robust Detection of Binary Encoded Malware," arXiv, 2018. [Online]. Available: https://arxiv.org/pdf/1801.02950

[8] RAVIT DOTAN et al., "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework," arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2401.15229