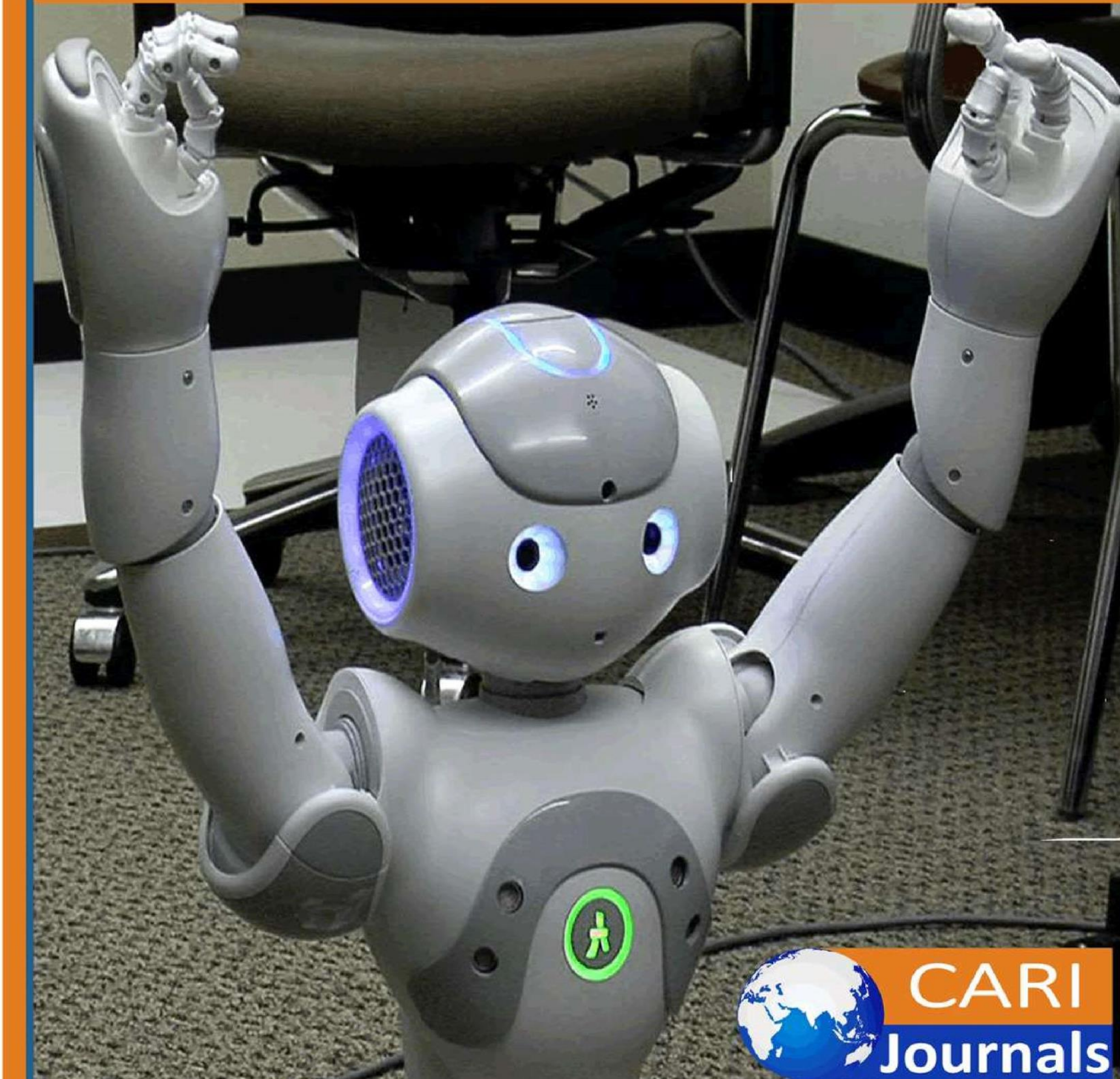
International Journal of
Computing and
Engineering
(IJCE)

AI Model Security and Adversarial Techniques in Finance and Accounting Analytics



CARI
Journals

AI Model Security and Adversarial Techniques in Finance and Accounting Analytics

Pratik Koshiya
https://orcid.org/0009-0003-1781-9266

Accepted: 11th July, 2025, Received in Revised Form: 18th July, 2025, Published: 25th July, 2025

Abstract:

AI integration in finance and accounting domain is revolutionizing core areas such as fraud detection, algorithmic trading, credit underwriting, and audit analytics. So, while they make operations more efficient and improve decision-making, and this introduce unique challenges and security threats specific to AI systems. This paper aims at giving a complete rundown of such vulnerabilities which are fast emerging for financial AI models, such as data poisoning, adversarial inputs, model extraction, and membership inference attacks, and provides examples based on real-world financial scenarios to show how these attacks could counter fraud detection, influence market behavior, or compromise sensitive data. The study also considers practical defense mechanisms-including adversarial training, input validation, privacy-preserving methods, and live model monitoring-that could be instituted at any point along the AI value chain. Recognizing the urgent need of robustness and regulatory compliance, the paper advocates “security-by-design” approach facilitated by cross-functional teams. These insights are intended to help both practitioners and policy makers work towards the creation of secure and trustworthy AI systems that meet operational and regulatory demands required by the modern financial ecosystem.

Keyword: *AI Model Security, Adversarial Techniques, Finance analytics, Accounting Analytics*

Introduction to AI in Finance and Accounting & The Importance of Security

The Transformative Role of AI in Finance and Accounting

Artificial Intelligence (AI) and Machine Learning (ML) are rapidly reshaping the finance and accounting landscapes. These technologies are no longer futuristic concepts but are integral to daily operations, strategic decision-making, and innovation within the financial sector. Key applications such as Fraud Detection and Prevention where AI algorithms excel at identifying anomalous patterns in vast datasets of transactions, flagging potentially fraudulent activities in real-time for credit cards, insurance claims, and payments. In Algorithmic Trading, ML models analyze market data, news sentiment, and economic indicators to execute trades at high speeds, aiming to optimize returns and manage risk. Credit Scoring and Loan Underwriting applications where AI provides more nuanced risk assessments by analyzing a wider range of data points beyond traditional credit reports, leading to more accurate and inclusive lending decisions. For Risk Management scenarios predictive models are used to forecast market volatility, credit default risks, operational risks, and ensure regulatory compliance (e.g., Anti-Money Laundering - AML, Know Your Customer - KYC). To perform Financial Forecasting and Advisory**,** AI tools assist in predicting market trends, company performance, and provide personalized financial advice to clients (robo-advisors). For Audit and Compliance Analytics, AI automates a significant portion of audit processes, such as reviewing large volumes of documents, identifying discrepancies, and ensuring adherence to accounting standards and regulations. Last but not least Customer Service domain, AI-powered chatbots and virtual assistants handle customer inquiries, provide support, and guide users through financial processes. The benefits are substantial, ranging from increased efficiency and reduced operational costs to improved accuracy in predictions, enhanced customer experiences, and the ability to uncover insights previously hidden in complex data.

The Emerging Threat Landscape: Why AI Model Security is Critical

As financial institutions increasingly rely on AI, these systems themselves become attractive targets for malicious actors. The very intelligence and automation that make AI powerful also introduce new vulnerabilities. Attacks on AI in finance can cause direct losses: Illegal interference with a target trading algorithm could cause disastrous trades, and fraudulent interference with a fraud detection mechanism could allow fraudulent transactions to go undetected. Apart from the financial imposition, the breaches may cause catastrophic damage to reputation, customer confidence and lowering business valuation. Subsequently, one can be levied with regulatory fines if sensitive data is misused or decision-making systems fail to meet a compliance standard, which might represent a large fine. In the worst cases, a coordinated attack could threaten destabilization of the financial markets. Once the model has been compromised, its predictions are no longer reliable, thereby defeating the purpose they exist for. Traditional cybersecurity measures, while still necessary, are often insufficient to protect AI models because the attack vectors are different. Adversaries can exploit the learning process, the data, or the model's decision-making logic itself. Therefore, a specialized focus on AI model security is paramount.

Understanding AI Model Vulnerabilities in Financial Contexts

AI models, particularly those based on machine learning, are not infallible black boxes. They possess inherent vulnerabilities that can be exploited. Understanding these vulnerabilities is the first step towards building secure AI systems in the high-stakes environment of finance and accounting.

Common Vulnerabilities in Machine Learning Models

Several classes of vulnerabilities are common across various ML model types. **Data Poisoning** occurs when an attacker intentionally injects corrupted or misleading data into the training dataset. The goal is to compromise the learning process, leading the model to learn incorrect patterns, make systematic errors, or create backdoors that the attacker can later exploit. For example, an attacker might feed a fraud detection system with subtly altered transaction data that labels fraudulent activities as legitimate, thereby teaching the model to ignore such fraud in the future. In algorithmic trading, poisoned historical price data could lead to flawed trading strategies.

Evasion Attacks are test-time attacks where an attacker crafts malicious inputs, often by adding imperceptible perturbations to legitimate inputs, to cause the model to misclassify them. The model itself remains unchanged. Such as, a loan applicant might slightly alter their financial details in a way that is difficult for a human to spot but fools an AI credit scoring model into granting a loan they don't qualify for. Fraudsters could subtly modify transaction details to bypass AML systems. **Model Inversion** attacks aim to reconstruct parts of the training data or sensitive features used to train the model by querying the model and observing its outputs. For example, If a model predicts an individual's risk of defaulting on a loan, a model inversion attack might attempt to infer sensitive personal financial information (e.g., income, specific debts) that was part of the training data, leading to severe privacy breaches.

Membership Inference attacks attempt to determine whether a specific data record was part of the model's training set. In simple scenario, this could reveal sensitive information, such as whether a particular individual's financial profile was used to train a model for a specific financial product or risk category, which could be a privacy violation or used for targeted attacks. **Model Stealing** aim to replicate or "steal" a proprietary, high-performance model by repeatedly querying it with various inputs and observing the outputs. They then use this information to train a clone model. A competitor might try to steal a successful algorithmic trading strategy or a sophisticated credit risk assessment model developed by another institution, thereby eroding competitive advantage. Backdoor Attacks is type of poisoning attack where the model performs normally on most inputs but behaves maliciously when a specific, attacker-chosen trigger (a subtle pattern or feature) is present in the input. For example, an attacker could insert a backdoor into an AML system such that transactions from a specific (illicit) source are always flagged as legitimate if they contain a hidden trigger.

Unique Challenges in Finance and Accounting

Securing AI models in the financial domain presents unique challenges due to sensitivity, complexity and regularity context of financial data. Financial data is highly sensitive (PII, transaction details, account information). Breaches can lead to identity theft, financial fraud, and severe reputational damage. Regulations like GDPR, CCPA, and others impose strict requirements. Additionally, the financial industry is heavily regulated. AI models used for critical functions like credit decisions, fraud detection, and compliance reporting must be robust, fair, explainable, and secure to meet regulatory scrutiny (e.g., SR 11-7 in the US for model risk management). The stakes are high, as successful attacks can result in immediate and significant financial losses. Compounding these risks is the complexity of financial data, which is typically high-dimensional, time-series based, and non-stationary, posing unique modeling and security challenges. There is a critical need of Explainability (XAI). Regulators and internal governance often require that AI decisions, especially adverse ones (e.g., loan denial), be explainable. However, some techniques to make models more robust can sometimes reduce their interpretability, creating a trade-off. At last, Adversaries are constantly developing new techniques, requiring continuous adaptation of defense mechanisms. Addressing these vulnerabilities requires a holistic approach that integrates security considerations throughout the AI model lifecycle, from data acquisition and preparation to model development, deployment, and ongoing monitoring.

Adversarial Techniques Targeting Financial and Accounting AI

Adversarial attacks on AI systems are sophisticated techniques designed to deceive machine learning models. In the context of finance and accounting, these attacks can be tailored to exploit specific applications, leading to significant harm. Understanding the mechanics of these attacks is crucial for developing effective defenses.

Evasion attacks are the most common form of adversarial attack also known as test-time attack. The attacker's goal is to craft an input that the model misclassified at the time of prediction (test time), without altering the trained model itself. An attacker makes small, often human-imperceptible, modifications to a legitimate input sample. This modified sample, known as an adversarial example, is then fed to the target model, causing it to produce an incorrect output desired by the attacker. In generating adversarial examples, the idea is to force an incorrect response from the model. In finance and accounting, evasion attacks can be in any form: a loan applicant with poor credit will make subtle alterations so as to appear low-risk, particularly adjusting his income or debt-to-income realization measurements; A fraudster could modify transaction details (e.g., merchant category code, transaction amount, IP address) in a way that makes a fraudulent transaction appear legitimate to an AI-based fraud detection system; Malicious actors could structure financial transactions or alter identifying information to evade detection by AI systems designed to flag suspicious activities related to money laundering; or Perturbing input data such as news sentiment scores or micro-price movements and fed to a trading algorithm to manipulate model. Common techniques may be used in here such as Fast Gradient Sign Method (FGSM), A one-step method that adds a small perturbation in the direction of the gradient of the loss function with respect to the input. It's fast but often less effective than iterative methods., Projected Gradient Descent (PGD) which is an iterative version of FGSM, where small steps are taken in the gradient direction, and the result is projected back onto a permissible input region, Carlini & Wagner (C&W) Attacks, which are powerful, optimization-based attacks that are often more effective at generating adversarial examples, though computationally more expensive; and the Jacobian-based Saliency Map Attack (JSMA) Focuses on modifying a minimal set of input features that have the most impact on the model's output.

Poisoning attacks which also knowns as training-time attack, aim to compromise an AI model during its training phase by inserting malicious or misleading data into the training set. Such attacks affect the learning of the model, convincing it to adopt incorrect decision boundaries so that it erroneously classifies certain inputs or simply has degraded performance in general. The financial and accounting domain can be severely affected. In fraud detection systems, for example, an attacker could slowly inject transactions that are fraudulent but labeled as legitimate (or vice-versa) into the training data of a fraud detection system. Over time, this would teach the model to misclassify similar fraudulent transactions. Similarly, if an algorithmic trading model is trained on historical market data and news sentiment, an attacker could inject falsified historical data or skewed sentiment data to bias the model's future predictions for certain stocks: corrupt the financial records in training data, so that the model learns to ignore irregularities. A more sophisticated poisoning attack can create a "backdoor" or "Trojan" in the model. The model behaves normally for most inputs but misbehaves in a way desired by the attacker when a specific, subtle trigger (e.g., a particular sequence of characters in a transaction description, a tiny watermark in an image of a document) is present in the input. Typical strategies for poisoning attacks include label flipping, the attacker changes the labels of a subset of training samples, and data injection, where the attacker inserts new, crafted data points into the training set. These points can be designed to subtly shift decision boundaries or create specific vulnerabilities. Model Stealing or Extraction Attacks aim to create a functional copy of a proprietary (victim) AI model without direct access to its architecture or training data. Here, the attacker queries the victim model (often a publicly accessible API or a model embedded in a product) with a large number of diverse inputs and observes the corresponding outputs (predictions or confidence scores). This input-output behavior is then used to train a "clone" or "surrogate" model that mimics the functionality of the victim model. A competitor could attempt to steal a successful proprietary trading algorithm by observing its responses to various market scenarios (if inferable through an API or service). If a financial institution offers a service that provides credit risk assessments via an API, an attacker could query this API extensively to build their own version of the underlying model. Stolen models can lead to loss of intellectual property and competitive edge, as the effort and resources invested in developing the original model are effectively nullified.

Membership Inference Attacks try to determine whether a specific data record was part of the model's training data. By observing a model's prediction confidence or behavior on a given data point, an attacker can infer if that point was likely seen during training. Models sometimes "overfit" or memorize parts of their training data, leading to different responses for seen vs. unseen data. There are chanced of privacy breach which confirms that an individual's specific financial profile (e.g., a high-net-worth individual's investment patterns) was used to train a particular financial advisory model can be a serious privacy violation. Additional adversary attempts to identify data from a specific sensitive group (e.g., individuals who defaulted on loans) which was used in training, potentially for discriminatory purposes or targeted attacks.

Model Inversion / Attribute Inference Attacks aim to reconstruct sensitive features from the training data or even entire training samples by leveraging the model's outputs. Given a model and some partial information (e.g., a class label or some non-sensitive features of a data point), the attacker tries to infer the missing sensitive attributes or reconstruct a representative sample of that class. In financial context, if a model predicts a loan applicant's eligibility, an attacker might try to reconstruct sensitive inputs like exact income, specific debt amounts, or investment details that contributed to that prediction. For models trained on aggregated or anonymized corporate financial data, an attacker might attempt to infer specific sensitive details about individual companies. These adversarial techniques highlight the diverse ways AI models in finance and accounting can be compromised. The financial sector must be acutely aware of these threats to design and deploy AI systems that are not only accurate and efficient but also resilient and secure.

Defense Mechanisms and Mitigation Strategies

Given the array of potential attacks, securing AI models in finance and accounting requires a multi-layered defense strategy. No single solution is foolproof; instead, a combination of techniques applied throughout the AI lifecycle offers the best protection. One major way is to bolster training datasets with adversarial examples through adversarial training to make a resilient model. This strategy can help fraud detection systems to correctly classify altered malicious transactions. Defensive distillation constitutes another technique whereby outputs are softened by a teacher model to train a student model, this process can smooth the model's decision surface, making it harder for attackers to find exploitable gradients. Regularization techniques such as adding terms to the loss function during training that penalize model complexity (e.g., L1/L2 regularization, dropout). This can prevent overfitting and make models less sensitive to small input perturbations. Certified defenses go further by training methods that can provide a formal guarantee (a certificate) that the model's output will not change for any input within a certain bounded perturbation. Input sanitization techniques also play a crucial role as model like Anomaly Detection for Inputs, Feature Squeezing and Data Transformation which can mitigate adversarial perturbations before they reach model and treats inputs. Such scenario where Before a transaction is fed to a fraud detection model, an input anomaly detector could flag it if its characteristics are highly unusual or applying transformations (e.g., adding random noise, spatial smoothing for image-based data like scanned documents) to inputs to try and "undo" adversarial perturbations. The choice of model architecture itself can influence its inherent robustness. Some model architectures (e.g., random forests, gradient boosted trees, under certain configurations) can be less susceptible to certain types of adversarial attacks compared to, for example, very deep neural networks with highly complex decision boundaries, though this is an area of active research. Combining predictions from multiple diverse models. An attacker would need to fool a majority or a weighted combination of models, which can be more challenging. It Can improve both accuracy and robustness, but it may Increase computational cost. Proactively testing model’s security for vulnerabilities is crucial such as Red Teaming for AI Models, Vulnerability Scanning and Formal Verification Methods. Employing a dedicated team to simulate attacks on AI systems using known adversarial techniques to identify weaknesses before they are exploited externally. Using specialized tools and frameworks (e.g., Adversarial Robustness Toolbox (ART), CleverHans, Foolbox) to systematically test models against a battery of attacks. Mathematical techniques to prove certain properties about a model's behavior, such as its resistance to specific types of perturbations. (Often related to certified defenses). Finally, Protecting the data used to train and operate AI models is fundamental. Differential privacy and federated learning, secure multi-party computation and robust data sourcing and cleansing. This is vital to protect privacy and limit exposure to membership inference and model inversion. Implementing these defense mechanisms requires a collaborative effort between data scientists, machine learning engineers, cybersecurity professionals, and risk management teams within financial institutions.

The Future of AI Security in Finance and Accounting & Conclusion

The landscape of AI security in finance and accounting is dynamic, characterized by an ongoing interplay between evolving adversarial capabilities and advancing defensive strategies. As AI becomes more deeply embedded in critical financial infrastructure, the stakes for ensuring its security and trustworthiness will only continue to rise. Several key areas are shaping the future of AI security in this domain. Explainable AI (XAI) is often discussed for transparency and bias detection; it can also aid security. Understanding *why* a model makes a certain prediction can help identify if it's being influenced by adversarial perturbations or if its logic has been subtly altered by poisoning. Conversely, attackers might also try to exploit XAI to better understand model vulnerabilities. Future work will increasingly address robustness in more complex tasks prevalent in finance, such as sentiment analysis in trading, document analysis in audit, time-series forecasting, for trading or fraud detection strategies. Trusted Execution Environments - TEEs, AI-specific chips with built-in security features could provide more secure platforms for training and deploying AI models, protecting them from certain software-based attacks. Use of AI itself to detect attacks or recognize adversarial patten recognition for other AI systems Standardization of Testing and Benchmarking of AI model are being developed industry wide to compare different defense mechanisms and ensure a minimum level of resilience. Finally, recognizing that fully automated defenses may not always be sufficient, future systems will likely involve human oversight and intervention, where AI flags suspicious activities for review by financial analysts or security experts. The regulatory environment is catching up with the rapid adoption of AI. Several initiatives will significantly impact AI security in finance. NIST AI Risk Management Framework (RMF) provides a voluntary framework for organizations to manage risks associated with AI, including security. EU AI Act introduced comprehensive legislation categorizes AI systems by risk level, imposing stricter requirements (including security and robustness) for high-risk applications, many of which are found in finance (e.g., credit scoring, critical infrastructure). Financial regulators worldwide (e.g., central banks, securities commissions) are issuing their own guidelines and expectations for AI governance, model risk management, and security within the institutions they oversee. Regulations like GDPR and CCPA continue to emphasize the secure and ethical handling of personal data, which is foundational to training secure and unbiased AI models. Compliance with these evolving regulations will necessitate a proactive and well-documented approach to AI security. Ultimately, Technology and processes alone are not enough. Financial institutions need to foster a security-aware culture that extends to AI development and deployment. Training and Awareness for Educating data scientists, ML engineers, and business stakeholders about AI-specific threats and best practices. Breaking down silos between AI development teams, cybersecurity teams, risk management, legal, and compliance departments. A "secure-by-design" approach should be embedded to Integrate security considerations from the very beginning of the AI development lifecycle, rather than treating it as an afterthought.

Conclusion: The Imperative for Proactive AI Security

The integration of AI into finance and accounting offers immense potential for innovation, efficiency, and enhanced decision-making. However, this reliance also introduces a new frontier of security challenges. Adversarial actors are continuously refining their techniques to exploit vulnerabilities in AI systems, and the financial consequences of successful attacks can be devastating. The defense against such threats is not a one-time fix but an ongoing arms race. It requires a multi-layered strategy encompassing robust model training, vigilant input validation, secure data governance, continuous monitoring, and adherence to emerging regulatory standards. Financial institutions must move beyond a reactive posture and proactively invest in the research, tools, talent, and processes necessary to build and maintain trustworthy and resilient AI systems. The future stability and integrity of the financial ecosystem will increasingly depend on the ability to secure its intelligent core.

Reference:

[1] M. Korolov, "How AI can help you stay ahead of cybersecurity threats," *CSO Online*, Oct. 19, 2017.

[2] T. Sweeney, "8 ways to spot an insider threat," *Dark Reading*, Sept. 6, 2019.

[3] M. Korolov, "How AI can help your organization stay a step ahead of cyberattackers," *csoonline*, Oct 2017.

[4] A. Chakraborty, A. Biswas, and A. K. Khan, "Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation," *arXiv*, Sep. 2022.

[5] M. Schmitt, "Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence-enabled malware and intrusion detection," *arXiv*, Oct. 15, 2023.

[6] S. Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," *arXiv*, Dec. 30, 2023.

[7] I. H. Sarker, H. Janicke, L. Maglaras, and S. Camtepe, "Data-Driven Intelligence can Revolutionize Today's Cybersecurity World: A Position Paper," *arXiv*, Aug. 9, 2023.

[8] Federal Reserve Board, "SR 11-07: Guidance on Model Risk Management," Apr. 4, 2011.

[9] Palo Alto Networks, "What are adversarial attacks on AI/Machine Learning," *Cyberpedia*.

[10] "AI use cases in financial services," *SmartDev*, 2023.

[11] "The role of Artificial Intelligence and Robotic Process Automation (RPA) in fraud detection: Enhancing financial security through automation," *ResearchGate*, 2023.

[12] "NIST AI Risk Management Framework," *Wiz.io Academy*, 2024.

[13] "How OWASP guidelines secure your AI systems," *Salesforce Blog*, 2024.

[14] "Risks of AI in banks & insurance companies," *Lumenova.ai Blog*, 2024.

[15] "Must-have AI security policies for enterprises: A detailed guide," *Qualys Blog*, Feb. 7, 2025.

[16] J. Tamene, "AI-Based RPA's Work Automation Operation to Respond to Hacking Threats Using Collected Threat Logs," *Applied Sciences*, vol. 14, no. 22, Art. no. 10217, Nov. 2022.

[17] J. Tamene, "Detecting and Preventing Data Poisoning Attacks

on AI Models," 2025 PhotonIcs & Electromagnetics Research Symposium, Abu Dhabi, UAE, 4-8 May

[18] "The Role of Artificial Intelligence and Robotic Process Automation (RPA) in Fraud Detection: Enhancing Financial Security through Automation," *ResearchGate*, 2025.