UDC 004

THE AI ARMS RACE IN CYBERSECURITY: COMBATING DEEPFAKES AND AUTONOMOUS ATTACKS

Baktygulova Laula

Satbayev University

Introduction. Cybersecurity has always been a high-stakes domain defined by the continuous struggle between innovation in offense and resilience in defense. Historically, this race involved human adversaries exploiting software flaws and human error. However, the advent of sophisticated Generative Artificial Intelligence (AI) and Machine Learning (ML) has fundamentally transformed the landscape, ushering in an era where automation and intelligence are the primary weapons. This transformation is not merely an upgrade to existing tools; it represents a paradigm shift where the scale, speed, and sophistication of cyber threats are now exponentially amplified.

The core challenge facing organizations today is the rise of the AI Arms Race: a cycle where advanced AI capabilities, once a defensive advantage, are rapidly being weaponized by malicious actors. This article asserts that while AI presents unprecedented challenges – from the creation of hyper-realistic deepfakes that shatter trust in digital communication to the deployment of autonomous and polymorphic malware that evades traditional detection – it simultaneously offers the only viable path for robust defense. To maintain security and resilience, organizations must understand the dual nature of AI, proactively deploy intelligent counter-measures, and establish a new model of Human-AI partnership to protect against the invisible, adaptive, and autonomous threats of the digital future.

Main body. The most immediate and concerning threat stems from deepfakes. Generative AI can synthesize highly convincing audio and video impersonations of key personnel – executives, colleagues, or trusted third parties – exploiting the weakest link in the security chain: the human element [1].

Voice Cloning and BEC Scams. Attackers utilize minimal voice samples (e.g., from voicemail or public videos) to generate real-time audio deepfakes. These are deployed in "CEO fraud" or Business Email Compromise (BEC) calls, where the attacker, impersonating an executive's voice, demands an urgent, unauthorized wire transfer. The psychological authenticity of the voice often bypasses a victim's critical thinking. Figure 1 shows an example of telephone fraud.

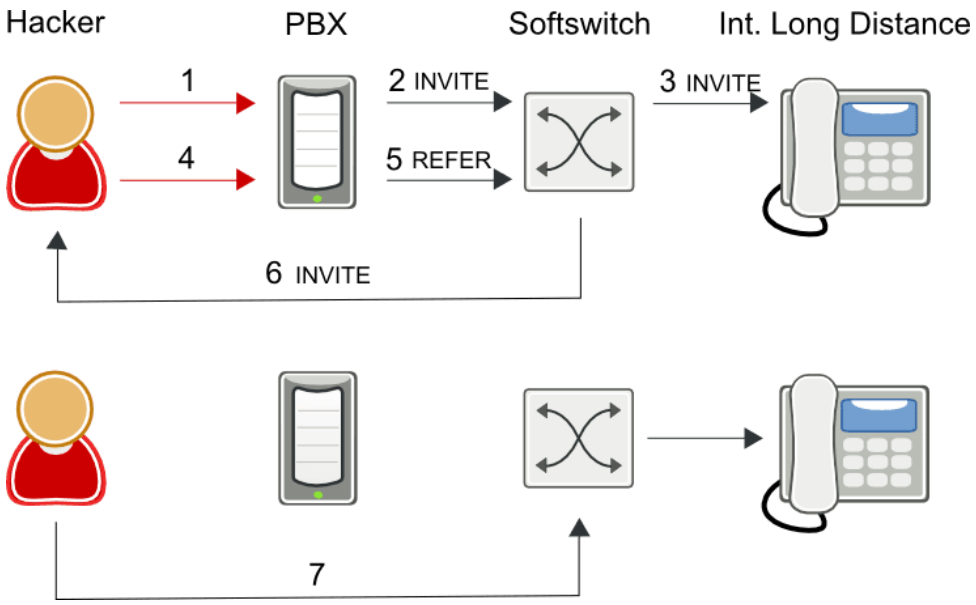

Figure 1. Telephone fraud scheme

Video Deepfakes in Targeted Attacks. While resource-intensive, high-fidelity video deepfakes are increasingly used in spear-phishing campaigns. A deepfake video of a "manager" requesting employees to install a "new mandatory meeting tool" lends unprecedented visual credibility to malicious requests, effectively eroding trust in digital communication protocols.

Autonomous and Polymorphic Malware. Traditional signature-based antivirus solutions struggle against the speed, volume, and versatility of AI-generated attacks [2]:

Automated Reconnaissance and Exploitation. AI can efficiently scan vast network segments, identify specific vulnerability chains, and dynamically tailor

exploits to target systems without continuous human intervention. This enables the execution of zero-day attacks or complex, multi-stage intrusions at machine speed.

Polymorphism and Evasion. Generative models can constantly rewrite and mutate their malicious code, making every instance of the malware unique. This polymorphic characteristic allows the malware to evade standard security sandboxes and static signature-based detection filters. An AI-controlled strain of malware can adapt its code in real-time to observed defensive measures, making it nearly impossible to track or block using conventional perimeter defenses.

Security teams are aggressively integrating AI into their defensive stack to regain the advantage of speed, scale, and proactive analysis.

Advanced Threat Detection (UEBA and NTA). AI and Machine Learning models excel at identifying subtle, complex patterns indicative of sophisticated threats, including advanced persistent threats (APTs) and "low and slow" attacks [3].

Behavioral Anomaly Detection (UEBA). Instead of relying on predefined "bad" lists (signatures), User and Entity Behavior Analytics (UEBA) utilizes AI to establish a baseline of "normal" behavior for every user and network entity. Any statistically significant deviation – such as a user accessing files at an unusual hour, logging in from a new geography, or exceeding typical data transfer volumes – triggers an immediate, risk-scored alert. This is crucial for catching insider threats and compromised accounts.

Automated Triage and Prioritization. In high-volume environments generating billions of log events daily, AI systems automatically correlate, score, and prioritize thousands of potential security alerts. This significantly reduces the Mean Time To Detect (MTTD) and Mean Time To Respond (MTTR) by allowing human analysts to focus exclusively on critical, high-fidelity incidents.

Securing the AI Models Themselves (AI GRC). A critical and cutting-edge defense is ensuring the integrity and trustworthiness of the very AI systems used for security. Threat actors can attempt to compromise or subtly manipulate

defensive models—an attack vector known as Data Poisoning or Adversarial Attacks.

AI GRC (Governance, Risk, and Compliance). Organizations must implement robust frameworks to audit, validate, and secure the massive datasets used to train their security models. This mitigates the risk of an attacker injecting malicious data points (e.g., labeling known malware samples as safe) into the training data, which could blind defensive AI to specific threats.

Adversarial Robustness. Defensive strategies must include continuous testing of models against adversarial attacks, where attackers introduce minimal, often imperceptible, changes to input data (like a single pixel change in an image) to force the AI model into a misclassification, ensuring the defensive AI remains reliable under stress [4].

Conclusion. The AI Arms Race dictates a future where cybersecurity is no longer merely a fixed set of rules, but a continuous, high-speed interaction between intelligent systems. The decisive advantage in this conflict will belong not to those who possess the most advanced AI, but to those who can effectively integrate human judgment and expertise with automated defense. The ultimate security posture in the AI era is defined by cyber-resilience. This necessitates the mandatory implementation of robust foundational protocols (such as Multi-Factor Authentication (MFA) and Zero Trust architectures), supplemented by sophisticated AI-driven threat detection. Ultimately, this battle is a partnership where AI provides the speed and scale while humans provide the context, strategy, and ethical oversight. Success in the AI arms race depends on our ability not just to react, but to proactively anticipate threats and rapidly adapt our intelligent defense systems.

Literature:

1. René Boiselle. (2025). Deepfakes Unmasked: Enterprise Cyber Security in the Age of AI Manipulation and Countermeasures. [Journal Article/Conference Paper].

2. Medium. Rules, Rules Everywhere: Why Signature-Based Detection Falls Short Against AI Threats. [Online]. Available: [https://medium.com/deeptempo/rules-rules-everywhere-why-signature-based-detection-falls-short-against-ai-threats-6bce79f7e974]
3. Lorven Technologies. (2025). AI Threat detection. [Online]. Available: [https://lorventech.com/ai-threat-detection/]
4. ACM Computing Surveys. (2024). Adversarial Attacks and Countermeasures on Image Classification-based Deep Learning Models in Autonomous Driving Systems: A Systematic Review [Online]. Available: [https://dl.acm.org/doi/full/10.1145/3691625]