# New cybersecurity threats emerge as businesses integrate AI models



As businesses increasingly integrate artificial intelligence (AI) models, particularly language large models (LLMs), into their operations, a new category of cybersecurity threats has emerged, which conventional security systems are ill-equipped to detect or manage. The rapid deployment of AI-based tools for customer assistance, internal support, and document automation exposes companies to direct attacks on the AI models themselves, their training data, and their interactive processes.

Traditional cybersecurity measures like firewalls, data loss prevention (DLP) systems, and proxies, primarily designed to guard endpoints, networks, and access points, lack the capacity to secure the internal interactions between AI models and their databases or to monitor the content AI systems accept and generate in real time. As a result, security operations centres (SOCs) often miss conversational anomalies, DLP solutions are not adapted for AI-specific data formats such as vector embeddings, and proxies fail to inspect requests directed at third-party LLM APIs.

Cybersecurity experts have identified three principal angles of attack targeting AI models:

1. **Data Poisoning:** This involves deliberate contamination of the model's training dataset to manipulate its behaviour. Such alterations can introduce biases, inject malicious content, or spread disinformation. In environments supporting continuous learning, these manipulations can propagate unnoticed, silently compromising the model's integrity.

2. **Prompt Injection:** Attackers insert hidden commands or content designed to bypass safeguards within user inputs. This technique can alter the model's responses or extract confidential information, effectively turning users into vectors for exploitation.

3. **LLM Compromise:** This category targets the AI infrastructure itself, including unauthorised access to API tokens, exploitation of window context vulnerabilities, abuse of service quotas, or the deployment of complex prompt chains engineered to induce unexpected or harmful model behaviour.

Given these novel threats, companies such as Palo Alto Networks have begun developing specialised Runtime Protection layers tailored specifically for AI models. The aim is to monitor, isolate, and regulate the interactions among AI models, users, plugins, and data sources. Three key components of this approach have been highlighted:

* LLM Gateway / Firewall: Implements filtering mechanisms for incoming queries, blocks injection patterns, and standardises inputs. Examples include solutions such as Promptshield, Promptarmor, and Protect AI Inference Shield.
* Data Flow Monitor: Inspects internal queries and communications between the model, databases, and plugins. Solutions cited include Palo Alto AI Runtime, Robust Intelligence, and HiddenLayer.
* API Access Governance: Controls and regulates access to open-source or third-party AI models, with providers such as Appomni, Cradlepoint, and Cloudflare API Shield offering relevant tools.

This evolving security landscape requires a significant shift in the posture of chief information officers (CIOs). AI models differ fundamentally from traditional applications; they are scalable, operate independently, and often interact with sensitive data. Thus, protection strategies must move beyond perimeter defence towards securing cognitive interactions. Key measures include mapping deployed AI models and their dependencies, analysing prompt behaviours for anomalies, and the capacity to isolate compromised models in real time.

In this AI-first environment, the attack surface extends from system infrastructure to the very logic and inputs governing AI behaviour. What users communicate with AI can become vulnerabilities. Ignoring this new dimension of security could leave companies with incomplete protection of their digital assets. The insights are detailed in a report from businesslife.co, which explores how enterprises must adapt their cybersecurity frameworks to address these emerging threats in the AI era.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.darktrace.com/blog/why-artificial-intelligence-is-the-future-of-cybersecurity> - This URL explains how AI's integration into cybersecurity introduces new security risks, including the misuse of AI by adversaries and attacks on AI models themselves. It highlights the challenges in detecting novel social engineering attacks fueled by AI.
2. <https://www.morganstanley.com/articles/ai-cybersecurity-new-era> - This article discusses the evolving landscape of AI and cybersecurity, noting that both cybersecurity teams and hackers are leveraging AI, creating a new era of threats and security measures.
3. <https://news.microsoft.com/source/canada/features/ai/how-ai-is-transforming-cybersecurity-tackling-the-surge-in-cyber-threats/> - This article showcases Microsoft's use of AI in processing vast amounts of signals to identify threats, highlighting AI's role in transforming cybersecurity amidst rising threats.
4. <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-cybersecurity-providers-next-opportunity-making-ai-safer> - The article discusses the challenges and opportunities AI presents for cybersecurity providers, including the need to protect AI itself from threats and the expanding attack surface due to AI integration.
5. <https://www.cobalt.io/blog/top-40-ai-cybersecurity-statistics> - This blog lists various statistics on AI cybersecurity, showing how AI is increasingly used in cyberattacks, including phishing and ransomware, which aligns with the evolving threat landscape described.
6. <https://www.darkreading.com/ai-machine-learning/why-cybersecurity-needs-ai-to-stay-ahead-of-threats> - Though not directly linked to AI model attacks, this article supports the broader idea that conventional cybersecurity measures are insufficient against new AI-driven threats and highlights the need for advanced AI-based solutions.
7. <https://news.google.com/rss/articles/CBMivwFBVV95cUxOZnREdUJNaGVhZEo2QV9oSEJMQm9aVWRncXlHWllkdVJiSkNIUEI4REtnMDJYS2ZNckttWUlOallDb0F1a1g4ckdzNWl5WEZSX1dMc2pKVUc1amJhcWxoVV94cmRvY1pZM1FrQU52TnA1TFlNSTZIb09LSE03Mk9TVXhXUzFJVzV6QVJVeEtsUjM2akpQMkNXZThqQWk3MC16VE91a3N1WmVySEZCVkY5LXlCT0VvY3RPanA4TlZnSQ?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data