# Ensuring AI security amidst rapid technological advancements



As the integration of artificial intelligence (AI) into business processes accelerates, organisations are facing an urgent challenge to ensure that their security frameworks can keep pace with technological advancements. With momentum building towards 2025, the need to secure AI deployments has never been more pressing.

Historically, discussions around AI security have been predominantly concerned with external threats, primarily focusing on safeguarding against AI-powered cyberattacks. However, a critical aspect that has not received adequate attention is the internal mechanisms of AI systems, particularly the ‘hidden layers’ within machine learning models. These layers, which lie between input data and output predictions, are vital for AI's ability to identify complex patterns and deliver nuanced results tailored to specific tasks. Yet, their integral role also renders them susceptible to manipulation by malicious entities. As Michael Adjei, Director of Systems Engineer at Illumio, noted, “the hidden layers of AI operate through learned representations, which can be difficult to interpret and monitor from a security point of view.”

One of the dominant concerns is the phenomenon of adversarial attacks. These attacks can subtly alter input data—changes often imperceptible to humans—but they have the potential to lead AI systems to produce erroneous or even harmful outputs. This vulnerability primarily lies within the hidden layers, where decisions are made based on learned patterns. The implications of such manipulation can be dire, particularly for high-stakes industries like healthcare and finance, where a misstep could lead to significant crisis scenarios.

The security challenges are further compounded by vulnerabilities present in the AI supply chain, which encompasses various elements such as data sources, training environments, software libraries, and hardware components. Compromise of any single component within this chain can jeopardise the security of the entire AI system. For example, if an adversary gains access to datasets for training AI models, they could introduce corrupt data, creating inherent biases or instabilities. This risk is particularly acute in sectors where accurate outputs are paramount, such as autonomous vehicles and financial advisory systems.

Another avenue of risk arises from the reliance on third-party AI services and platforms. As companies increasingly integrate pre-trained models and open-source libraries, they expose themselves to potential vulnerabilities stemming from these external resources. The security of such systems can be undermined by backdoors or security flaws embedded within third-party tools. To mitigate these risks, organisations are urged to adopt a Zero Trust security framework, which limits access on a need-to-know basis. This principle is aligned with the idea that AI models should not be trusted by default.

Isolation of AI workloads and rigid access controls are essential strategies within this framework to mitigate potential security issues. By carefully managing the interactions of AI systems with other data and systems, the attack surface may be minimised, thus protecting the integrity of AI deployments.

In terms of proactive measures, techniques like adversarial training can equip AI models to withstand manipulation. By introducing adversarial examples during the training phase, models can learn to detect and resist potential threats. Additionally, enhancing model interpretability tools will enable better understanding of how decisions are made, thereby improving the detection of security breaches.

Furthermore, there is a move towards establishing secure frameworks that govern the development and deployment of AI models. These frameworks aim to provide developers with robust tools designed to lessen vulnerability to attacks while incorporating mechanisms that can flag unusual behaviours or outputs from AI systems.

Looking ahead, a collaborative approach is deemed essential to forging stronger AI security protocols. Given the multifaceted nature of the risks associated with AI, stakeholders from various industries must cooperate to establish comprehensive security standards and regulatory measures that ensure AI systems remain secure and reliable.

Individually, organisations are also encouraged to embrace a Zero Trust mindset, incorporating principles such as "assume breach," least-privilege access, and “never trust, always verify.” Adopting a Zero Trust architecture can safeguard AI models against vulnerabilities, allowing businesses to harness AI’s transformative potential while maintaining robust security measures.

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

## Bibliography

1. <https://www.jamf.com/blog/jamf-insights-2025-ai-security/> - Corroborates the need for enhanced security measures in AI deployments, particularly highlighting the risks associated with AI-powered attacks and the importance of securing Apple devices.
2. <https://www.govinfosecurity.com/ai-risks-cybersecurity-challenges-for-2025-a-27212> - Supports the discussion on AI security challenges, including the vulnerability of hidden layers in machine learning models and the risks of adversarial attacks.
3. <https://www.govinfosecurity.com/ai-risks-cybersecurity-challenges-for-2025-a-27212> - Details the security concerns related to the AI supply chain, such as compromised data sources and training environments.
4. <https://www.jamf.com/blog/jamf-insights-2025-ai-security/> - Highlights the risks associated with relying on third-party AI services and platforms, and the need for a Zero Trust security framework.
5. <https://www.govinfosecurity.com/ai-risks-cybersecurity-challenges-for-2025-a-27212> - Emphasizes the importance of isolation of AI workloads and rigid access controls to mitigate potential security issues.
6. <https://www.jamf.com/blog/jamf-insights-2025-ai-security/> - Discusses the proactive measures such as adversarial training to equip AI models to withstand manipulation.
7. <https://www.govinfosecurity.com/ai-risks-cybersecurity-challenges-for-2025-a-27212> - Supports the need for enhancing model interpretability tools to improve the detection of security breaches.
8. <https://www.jamf.com/blog/jamf-insights-2025-ai-security/> - Highlights the move towards establishing secure frameworks that govern the development and deployment of AI models.
9. <https://www.govinfosecurity.com/ai-risks-cybersecurity-challenges-for-2025-a-27212> - Emphasizes the importance of a collaborative approach to forge stronger AI security protocols and regulatory measures.
10. <https://www.jamf.com/blog/jamf-insights-2025-ai-security/> - Encourages organisations to adopt a Zero Trust mindset, including principles such as 'assume breach,' least-privilege access, and 'never trust, always verify.'
11. <https://betanews.com/2025/01/09/how-to-unlock-innovation-safely-in-the-ai-revolution/> - Please view link - unable to able to access data