Responsible AI

A New Executive Framework from MIT Sloan - Securing AI Systems

Securing AI Systems with MIT Sloan's Secure-by-Design Framework

Why Secure AI is Critical?

- AI brings customer experience, efficiency, and risk management advantages.
- Growing concern: Security risks are often overlooked in the rush to implement AI.
- “Few are thinking about the security risks that come with it from day one.”

Unique AI Security Challenges

Not just another IT system: AI’s reliance on data, continuous learning, and probabilistic results creates new threats.

Major threats include:

- Evasion & poisoning attacks (manipulated data)
- Model theft & inversion (stolen or reconstructed models)
- Prompt injection (malicious user input forces leaks/actions)
- Privacy attacks (sensitive data exposure)
- Hallucinations (systems confidently providing false answers)

Why Existing Standards Fall Short

- NIST and ISO security standards address only parts of the AI challenge.
- Traditional frameworks rarely sit at the intersection of AI, security, and design

What is NIST?

NIST and ISO security standards address only parts of the AI security challenge because they were originally designed for broader IT and cybersecurity risk, not the unique vulnerabilities of AI systems:

- NIST (National Institute of Standards and Technology): The NIST AI Risk Management Framework (AI RMF) provides valuable guidance for identifying and managing AI-related risks—including adversarial attacks, data poisoning, privacy concerns, and bias. However, it has key limitations:
  - It is voluntary with no enforcement mechanisms, so adoption depends on organizational commitment.
  - Translating its principles into actionable steps is challenging, especially for organizations with limited resources or expertise.
  - The framework is evolving, with best practices for practical implementation still emerging.
  - NIST itself acknowledges significant challenges in mitigating attacks on AI/ML systems, citing insufficient defenses against data and model manipulation as well as a lack of reliable benchmarks for evaluating these mitigations.
  - Some guidelines place a heavy burden on model developers, potentially neglecting risk management for downstream deployers and users—a gap in shared responsibility.

https://www.nist.gov/itl/ai-risk-management-framework

What is ISO?

- ISO (International Organization for Standardization): While ISO offers standards that cover data privacy, information security management (e.g., ISO/IEC 27001), and general technology risk, these standards do not specifically address AI's:
  - Susceptibility to adversarial attacks (like prompt injection and data poisoning)
  - Issues like model inversion, unique privacy leaks, or AI-specific performance failures
  - Continuous learning and dynamic behavior, which introduce novel risks absent from static IT systems.

- In summary, NIST and ISO standards are essential foundations but not yet fully equipped to address all AI-specific risks like adversarial manipulation, model theft, prompt attacks, and emergent behaviors; organizations must supplement them with AI-tailored frameworks and practices to achieve comprehensive security coverage.

https://www.iso.org/standard/42001.html
https://www.iso.org/standard/23894.html

The MIT Sloan Secure-by-Design Framework

Goal: Enable executives to ask the right questions early and integrate security from the ground up.

- 10 strategic questions, organized by core categories of AI project design.

The 10 Strategic Questions

Category	Key Question
Strategic Alignment	How can AI initiatives align with goals, budget, values, ethics?
Risk Management	How will we identify, assess, and prioritize AI-specific risks?
Control Implementation	What technical & process controls will address these risks?
Policy/Procedures	What policies ensure data quality, privacy & ethics?
Governance	What structure oversees AI lifecycle security & operations?
Technical Feasibility	Does the AI architecture fit our current infrastructure?
Resource Allocation	How much security effort is needed; how will we resource it?
Performance & Security Monitoring	What metrics track AI effectiveness and security?
Continuous Improvement	How will we adapt and evolve AI practices?
Stakeholder Engagement	How will we build shared responsibility for AI security?

Real-World Impact – C6 Bank Case Study

C6 Bank adopted the framework to manage risk in customer service, fraud detection, and operational AI.

Key benefits:

- Identified 19 critical design factors
- Developed a 4-part platform to safely separate experimental and production AI
- Built tailored governance and compliance tools

Results and Lessons

- Early, structured questioning revealed security blind spots.
- New policies, best practices, and stakeholder confidence were achieved.
- **“The most powerful thing about these 10 questions is they force you to think ahead.”**

Takeaways for Leaders

- Building secure AI requires a new, proactive approach—not retrofitting security.
- The Secure-by-Design Framework offers a practical roadmap to address emerging threats.
- Asking smarter questions at the start = stronger, more resilient AI systems.

Get Started

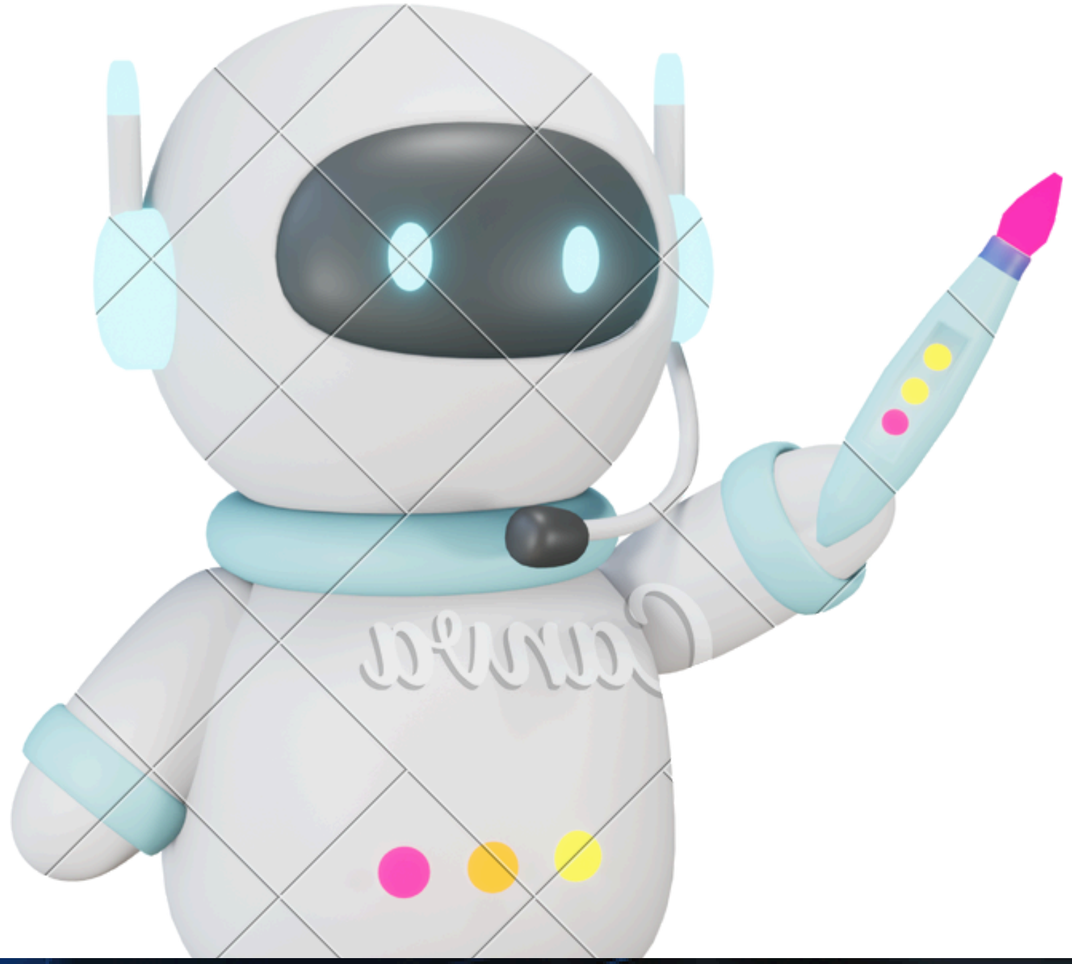
- Integrate the MIT Sloan 10-question framework into your AI initiatives today.
- Foster a culture of security, transparency, and continual improvement for AI innovation

Reference

- https://mitsloan.mit.edu/ideas-made-to-matter/new-framework-helps-companies-build-secure-ai-systems?utm_source=thinkingforward&utm_medium=email&utm_campaign=2025_07_29

Further Readings

- https://www.federalregister.gov/documents/2025/01/06/2024-30983/hipaa-security-rule-to-strengthen-the-cybersecurity-of-electronic-protected-health-information
- https://www.isaca.org/resources/glossary
- https://www.secondsight-ts.com/threat-assessment-blog/threat-and-risk-assessment-approaches-for-security
- https://www.cycognito.com/learn/exposure-management/security-controls.php



Thank You

Build Powerful AI.
Build with Ethics.
Build with Governance.
Build with Safety & Security
Build Together.