Trend Research

TREND MICRO

AI SECURITY STARTS HERE

The Do's and Don'ts Every Organization Must Know

Written by: **Dave McDuff** and **Andre Fernandes**

OCTOBER 14, 2025

TABLE OF CONTENTS

Introduction: Why AI Security Is Now a Board-Level Priority

The next wave of competitive advantage won't come from faster code or bigger datasets — it will come from AI systems companies can trust to operate safely, ethically, and compliantly.

From generative AI accelerating product design to autonomous AI agents executing business processes, the efficiency and innovation gains are enormous. But so are the risks:

- Manipulation via prompt injections
- Data leakage of regulated or proprietary information
- Poisoning of training or retrieval datasets
- Deepfake-based fraud and misinformation
- Supply chain vulnerabilities in pre-trained models and AI APIs
- Agentic Malicious SEO poisoning external data sources that AI systems rely on

In its "The Top Cybersecurity Threats in 2025" report, Forrester observed that "45% of DeepSeek's tests to generate harmful content bypassed safety protocols"[1] — revealing significant weaknesses in the model's safeguards. This highlights how easily attackers could weaponize less-secure AI models. Meanwhile, tightening regulations, such as the EU AI Act, China's Interim Measures for Generative AI Services, and various U.S. state-level AI laws, are raising the stakes for compliance, forcing organizations to be more careful about regulatory compliance, as violations could expose users and sensitive data to risks.

The message is clear: AI security is no longer optional. Building AI without embedded security is like constructing a skyscraper without a foundation. It might rise quickly, but it's dangerously unstable. Doing it right from "day zero" turns AI from a fragile experiment into a sustainable competitive advantage.

Business Case for AI Security

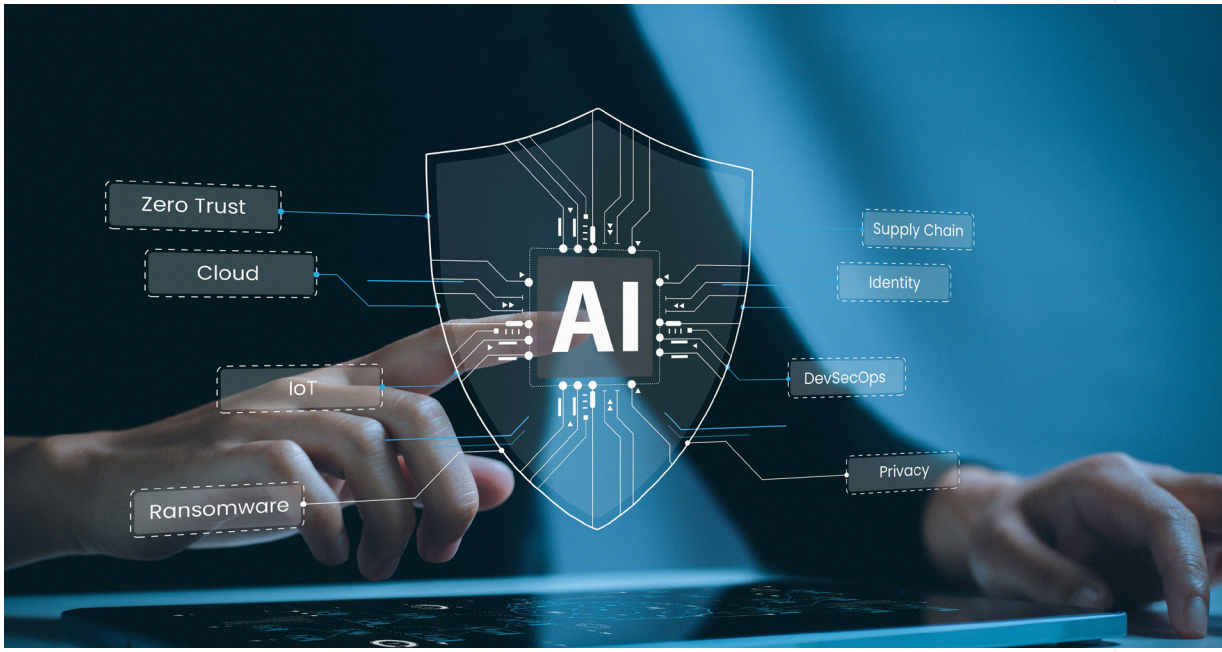
When organizations design security into the fabric of AI projects from the outset, they don't just avoid harm — they unlock a competitive advantage.

Accelerated Innovation, Lower Risk

- Integrating security from day one prevents costly reworks, breach responses, and compliance-driven delays. In DeepSeek testing, teams found that "78% bypassed safety protocols to create malicious or insecure code"[1] — a striking reminder that even cutting-edge systems can be vulnerable without rigorous security design. Testing models and applying appropriate guardrails early not only mitigates risk but also keeps projects on schedule and ahead of competitors.

TCO Reduction Through Prevention

- Security incidents in AI can instantly outweigh the cost of building the system itself. In 2024, "a deepfake 'CFO' successfully instructed a finance employee to transfer $25 million from the firm's bank account to a fraudulent account."[1] Preventive measures — such as adversarial testing, model monitoring, and user verification controls — cost far less than recovering from a single high-impact breach.



Customer Trust as a Differentiator

- ·Every successful AI product depends on user trust. Today, high-quality deepfakes in the form of manipulated images, synthetic voice cloning, and convincingly fabricated videos are already eroding verification systems for customers, partners, and employees alike. By implementing and showcasing transparent AI security practices, organizations can reassure stakeholders, ease adoption, and strengthen brand loyalty.

Regulatory Compliance Built-In

- Forrester's Business Risk Survey, 2024 found that "61% [of enterprise risk management decision-makers] expect their firm to spend more on regulatory compliance in 2025,"[1] underscoring the growing pressure organizations face to meet evolving laws and standards.
    - To manage this regulatory risk effectively, organizations can adopt AI cybersecurity frameworks such as NIST AI RMF, OWASP Top 10 for LLM Applications 2025, and CSA AI Control Matrix, which offer practical guidance on risk and control management.
    - SO/IEC 42001:2023, as an international standard, provides formal requirements for AI management systems and enables certification.
- By aligning these frameworks and standards, organizations are better equipped to comply with regulations such as the EU AI Act, reducing the likelihood of fines, forced product changes, or market restrictions — all while enabling AI initiatives to scale globally.

Future-Proofing

- AI threats are advancing rapidly and inexpensively. As Forrester notes, "Fraudsters can create interactive voice and video deepfake 'puppets' that they control for less than $5,000"[1] due to the proliferation of open-source algorithms, cheap GPU power, and readily available voice and audio profiles. An AI security platform with continuously updated threat intelligence can adapt quickly to these emerging risks, reducing the need for costly redesigns when attackers shift tactics.

AI Security by Design: The Do's and Don'ts

Domain	Do	Don't	Business Impact
Strategy and design	Integrate AI threat modeling, compliance mapping, and zero trust architecture from project inception	Build AI without oversight mechanisms or "kill switches"	Avoids rework and rogue AI usage
	Align with AI cybersecurity frameworks and develop the organization's usage and governance policies		
Supply chain security	Maintain a full Software Bill of Materials (SBOM) for models, datasets, packages, open-source libraries, and APIs to ensure transparency and traceability	Use unverified models or datasets without bias/ poisoning checks	Prevents poisoning, IP theft, and hidden vulnerabilities
Access and control	Apply least privilege + MFA	Give broad AI access to non-essential users or systems	Minimizes risk of malicious prompts or escalation
	Inspect prompts/responses at runtime		
	Monitor identities and AI agents for excessive agency.		
Operations and resilience	Ensure playbooks and red/blue teaming for top AI Threats	Ignore anomalies in embeddings, fail to patch AI dependencies	Blocks persistent attacks and retains forensic clarity
	Validate and sanitize all AI inputs, secure vector databases, log and review anomalies		
	Maintain model/data lineage		
People and governance	Train workforce on AI risk, deepfakes, data handling	Allow shadow AI, skip security awareness, ignore cultural misuse testing	Strengthens resilience to real-world misuse
	Mandate human-in-the-loop oversight		
	Include prank testing in red teaming		

People and Governance – The Human Dimension of AI Security

Technology alone can't secure AI. Human oversight, clear policies, and a culture of accountability are essential to prevent misuse, bias, and social engineering. As AI risks and regulatory pressures rise, organizations with strong governance and trained teams anticipate threats faster and maintain trust.

To translate that strategy into day-to-day practice, prioritize the following :

- **AI security training.** Ensure that staff members understand not only how to use AI tools but how they can be subverted.
- **Usage policies.** Codify what is allowed — and what is prohibited — in internal and customer-facing AI applications.
- **Human-in-the-loop oversight.** Any critical AI-driven output or action should have human review to meet safety and ethical standards.
- **Structured AI red teaming.** Beyond conventional pentests, simulate adversarial prompting, prompt injection chains, and data poisoning attempts.
- **Prank testing.** Simulate absurd or socially engineered misuse — e.g., ordering 1,000 tacos via drive-thru AI — to test operational resilience.

AI Frameworks and Compliance – Stay Ahead or Fall Behind

A compliance-first approach is critical as new frameworks, standards and regulations reshape the AI landscape. Embedding compliance from the start reduces legal and financial risks while building trust with customers and regulators. Organizations that prioritize compliance can scale AI securely and avoid costly disruptions.

Adopt the AI cybersecurity frameworks and standards and follow the regulations that affects the organization's jurisdiction:

Top AI Cybersecurity Frameworks and Standards

Framework/Standard	Description
OWASP Top 10 for LLM Applications 2025	Security risk list for large language models (e.g., prompt injection, sensitive data leaks)
NIST AI Risk Management Framework (AI RMF 1.0)	U.S. framework for managing AI risks, including security, bias, and resilience; widely adopted globally
ISO/IEC 42001:2023	First AI Management System standard; includes governance and security controls for AI lifecycles
Cloud Security Alliance AI Controls Matrix (CSA AICM)	Vendor-neutral framework with 243 AI security controls across 18 domains (e.g., model security, threat management)
MITRE ATLAS	Adversarial Threat Landscape for AI Systems; maps attack techniques and mitigations for AI models

Enacted AI Laws

Law or Act	Details
EU Artificial Intelligence Act (Europe)	A risk-based AI law that bans harmful uses (like social scoring), sets strict rules for high-risk AI systems, and requires transparency and human oversight
Transparency in Frontier AI Act (TFAIA) (USA - California)	Requires big AI developers to publish safety plans and report serious risks to prevent catastrophic AI failures
Colorado AI Act (CAIA) (USA – Colorado)	Makes companies using "high-risk" AI responsible for preventing bias and informing people when AI is used in decisions

Law or Act	Details
Section 103-E Artificial Intelligence (AI) Inventory (USA - New York)	Forces state agencies to list all AI tools they use and ensure workers' rights are protected when AI is involved
ELVIS Act or Ensuring Likness Voice and Image Security Act (USA - Tennessee)	Stops unauthorized cloning of someone's voice or image using AI and protects personal likeness rights
Interim Measures for Generative AI Services (China)	Regulates generative AI platforms by requiring content controls, security checks, and algorithm registration
Act 927 (GenAI ownership) (USA – Arkansas)	Clarifies who owns AI-generated content and enforces rules to prevent copyright violations by AI systems
Right to Compute Act (USA – Montana)	Protects lawful access to computing resources and sets rules for AI in critical infrastructure to manage risks

Note: Some jurisdictions (e.g., Singapore) have influential frameworks rather than binding AI statutes; those are listed in the Standards/Frameworks table.

Pending Legislation/Awaiting Enactment

Legislation Name (Country/State)	Proposed Coverage
Federal AI Safety Bill (USA)	Would require testing and reporting for advanced AI systems to prevent major risks
Texas Responsible AI Governance Act (USA – Texas)	Aims to ban manipulative AI practices and social scoring; sets transparency rules
UK AI Regulation Framework (United Kingdom)	Guides regulators to apply safety and fairness principles across sectors
Brazil AI Bill (PL 2338/2023) (Brazil)	Focuses on ethical AI and consumer protection; approved by Senate, awaiting final passage

Future-Proofing Against AI Threats

Emerging AI threats are accelerating in complexity and impact, spanning attacks on model integrity, misuse of generative capabilities, and unsanctioned adoption that can expose sensitive data and create legal or financial liabilities.

Emerging AI Threats	Description	Risk	Example
Indirect prompt injection	Attackers hide malicious instructions in trusted sources like emails, documents, or web pages, tricking AI systems into bypassing guardrails, invoking tools, and leaking or moving data.	In enterprise LLMs and agentic systems, these attacks are severe because they piggyback on trusted content and can execute silently without user clicks.	AIM Security's EchoLeak exposed a zero-click flaw in Microsoft 365 Copilot where a single crafted email triggered indirect prompt injection to exfiltrate sensitive data.
Poisoned training data	Attackers can manipulate even a tiny fraction of a model's training dataset to implant hidden behaviors or backdoors.	Causes models to misbehave when specific strings appear in prompts, enabling denial-of-service or data exfiltration attacks	Carnegie Mellon's CyLab researchers demonstrated that altering just 0.1% of a pre-training dataset can compromise an AI model. The poisoned data allowed attackers to embed backdoors that activate under certain conditions, proving how minimal changes can lead to major security risks.
Deepfake voice and video fraud	Attackers use AI-generated audio and video are used to convincingly impersonate individuals during calls or meetings, enabling social engineering and identity deception.	Insider access, data theft, malware on company devices, and legal exposure if hiring sanctioned actors	In July 2025, The US Justice Department dismantled a North Korea scheme in which fake IT workers used deepfake interviews to secure remote jobs and funnel earnings to the regime.

Emerging AI Threats	Description	Risk	Example
Model theft/ extractive attacks	Adversaries steal or replicate model weights, distill model behavior via APIs, or extract memorized training data from deployed systems.	Loss of IP and competitive advantage, privacy breaches from recovered training data, and compliance or legal exposure	In 2023, Meta's LLaMA model weights leaked online via BitTorrent, placing full model artifacts into public circulation.
Shadow AI adoption	Employees or teams use unsanctioned AI tools or deploy chatbots without governance, exposing sensitive data and bypassing security controls.	Data leakage, compliance violations, IP loss, and reputational or legal consequences	Healthcare workers uploaded patient data to AI tools and personal cloud accounts, creating HIPAA compliance breaches.

This is a list of best practices to use to protect against those emerging threats:

1. **Continuous threat intel**
   Monitor MITRE ATLAS, OWASP Top 10 for LLM Applications 2025, and industry AI threat groups. Convert new findings into actionable tickets.
2. **Dynamic governance**
   Perform quarterly reviews of AI inventory, risks, and policies aligned with ISO/IEC 42001:2023 and NIST AI RMF. Require owners, kill switches, and approved tool lists.
3. **Guardrail and prompt injection defense**
   Perform pre/post prompt filtering, least-privilege tool access, and adversarial testing for jailbreaks and indirect injections.
4. **Secure retrieval-augmented generation (RAG) and external content**
   Create an allowlist of sources, scrub HTML/links, enforce retrieval policies, and monitor vector store integrity.
5. **Data protection by design**
   Apply DLP and redaction pre-inference, scan outputs post-inference, and enforce policy-aware gateways.
6. **Supply-chain assurance**
   Demand signed artifacts for models/datasets, and screen for poisoning/ backdoors before deployment.

7. **Model theft mitigation**
   Rate-limit queries, watermark training data, segregate weights, and monitor for extraction patterns.

8. **Deepfake resilience**
   Require out-of-band verification for sensitive actions, and maintain a deepfake incident playbook.

9. **Shadow AI control**
   Publish a safe catalog of approved tools, block unsanctioned endpoints, and train staff on AI security.

10. **SOC integration**
    Deploy AI-powered platforms for automated incident response along with AI-driven playbooks. Additionally, embed AI threat hunting into SOC workflows: guardrail deletion, abnormal RAG calls, and refusal bypass attempts.

11. **Usage and cost telemetry**
    Track tokens, endpoints, and spend to detect unsanctioned use and exfiltration attempts.

Conclusion: AI Security as a Strategic Multiplier and Competitive Advantage

AI will define the next decade of business innovation. But without security, it can just as easily undermine trust, trigger regulatory breaches, and disrupt operations. The difference between an AI initiative that propels growth and one that stalls under scrutiny lies in how it is built — and whether security is treated as a strategic enabler from day zero.

Organizations that embrace secure, transparent, and well-governed AI will:

- **Innovate faster** by avoiding compliance bottlenecks and reducing project rework.
- **Lower total cost of ownership (TCO)** by preventing costly compromises before they happen.
- **Earn enduring trust** from customers, partners, and regulators.
- **Adapt fluidly** to new threats without the need for wholesale redesign.

By embedding robust safeguards into AI from the outset — hardening operational pipelines, securing supply chains, validating models, and continuously monitoring for emerging risks — organizations can unlock AI's full potential while protecting brand, revenue, and reputation.

How Trend Vision One Helps

Governance and Compliance

- **CREM** – Compliance Management: Enables organizations to assess, customize, monitor, and report their security posture against selected like NIST RMF or custom frameworks and standards
- **ZTSA AI Access Control:** MFA, prompt inspection
- **Endpoint Deepfake detection:** When enabled, conforms to compliance standards

Operational Resilience

- **AI Application Security:** Protects AI models and applications from vulnerabilities, malicious prompts, and data leaks using two capabilities. AI Scanner and AI Guard blocks
- **Container Security** and **Code Security:** Scan, patch, and protect AI workloads
- **File Security:** Scan datasets and artifacts for threats before training or deployment
- **Red and Purple Teaming Services:** Realistic attack simulations to strengthen detection and response

Threat Defense

- **AI Application Security** and **AI-DR:** Protect LLMs, RAG pipelines, and vector DBs
- **Deepfake Detector:** Stop synthetic identity and voice/video fraud
- **TippingPoint IDS/IPS:** Block AI-powered exploits at the perimeter
- **AI App Guard:** Specialized protection for AI applications and their associated files on user workstations
- **Trend Cybertron:** Predict and prioritize AI threats including cyberattack paths
- **Trend Companion AI** (Generative AI): Turn complex threat intel into executive-ready action plans

Ready to Get Started? Five Actions for This Topic

- ✓ Inventory all AI tools in use (sanctioned and shadow).
- ✓ Implement MFA on all AI system access.
- ✓ Schedule AI security training for development teams.
- ✓ Review and document your AI model supply chain (SBOM).
- ✓ Contact Trend Micro to start an AI risk assessment.

Reference

1. Allie Melen et al. (April 14, 2025). *Forrester.* "The Top Cybersecurity Threats In 2025." Accessed on Oct. 27, 2025, at https://www.forrester.com/report/the-top-cybersecurity-threats-in-2025/RES182329.

Want more insights like this?

TrendMicro.com/ai

AI SECURITY STARTS HERE

TREND MICRO™ | Proactive Security Starts Here

Trend Micro, a global cybersecurity leader, helps make the world safe for exchanging digital information between people, governments, and enterprises.

Trend leverages security expertise and AI to protect more than 500,000 enterprises and millions of individuals across clouds, networks, endpoints, and devices worldwide.

At the core is Trend Vision One™, the only AI-powered enterprise cybersecurity platform that centralizes cyber risk exposure management and security operations, delivering layered protection across on-premises, hybrid, and multi-cloud environments.

The unmatched threat intelligence delivered by Trend empowers organizations to proactively defend against hundreds of millions of threats every day.

Proactive security starts here. TrendMicro.com