Aristek

AI system under attack: the essentials of AI security

Like many new technologies, AI brings both opportunities and challenges in how we control and secure it.

Why? Because AI systems hold massive amounts of sensitive data.

Data is the lifeblood that fuels AI algorithms – and the prime target for cyberattacks.

Once breached, the impact can be serious:

- **In the short run** – financial losses, business interruptions, legal issues.
- **In the long run** – reputational damage and loss of customer trust.

82%
of C-suite executives say that secure AI is essential for the success of their company

24%
of current AI projects are secured properly

Source: IBM

So, how do we balance the opportunities AI offers with the security it demands?

Our CTO, Aleksei Turchak, has put together this go-to guide to help you understand and tackle the key aspects of AI security.

To understand **how** to safeguard AI, first know **what** is at risk

1 Data risks

Sensitive data faces multiple threats:

- **Poisoning** – adding false, incorrect, or irrelevant data to confuse the model
- **Exfiltration** – hackers access and steal critical datasets.
- **Leakage** – sensitive information is unintentionally exposed, compromising privacy.

One example of data leakage is when Microsoft researchers accidentally exposed **38TB** of private data, including employee backups.

2 Model risks

Attacks on AI models:

- **Stealing models** for unauthorized use
- **Reverse-engineer** to reveal proprietary algorithms
- **Manipulate architecture, weights, or parameters,** altering model behavior and accuracy.
- **Jailbreaking,** where attackers bypass restrictions to force models into unintended or harmful actions.

Did you know that vulnerabilities found in Hugging Face's platform could have allowed attackers to leak customer data?

3 Operational risks

AI systems are susceptible to:

- **DoS attacks,** crippling AI services and operations
- **Supply chain threats,** introducing vulnerabilities through third-party components
- **Model drift,** where changing data makes model accuracy over time.

In January 2024, cybersecurity researchers found critical vulnerabilities in SAP AI systems that exposed customer data and cloud environments.

4 Ethical / compliance risks

Risks affecting fairness and trust:

- **Bias,** causing unfair decisions
- **Lack of transparency,** making decisions hard to explain
- **Hallucinations,** where AI gives wrong or made-up answers.

A CVSS 10 bug in an LLM-to-SQL library showed how attackers could manipulate prompts to bypass system controls.

What about LLM (gen AI) and its risks?

Generative AI is one of the fastest-growing yet still maturing AI technologies. With its widespread adoption, it has become a focal point for cybercriminals.

Some risks of LLM overlap with general AI, but others are more unique.

Prompt injection
Attackers manipulate the model through direct inputs or hidden injections, altering its behavior.

Model theft
Unauthorized access and copying of proprietary models lead to security and IP risks.

Training data poisoning
Manipulated training data introduces vulnerabilities or biases into the model.

Model Denial of Service (DoS)
Resource-heavy attacks reduce AI system service quality for users.

Supply chain vulnerabilities
Weaknesses from outdated components, poisoned plugins, or poisoned data sources.

Sensitive information disclosure
Private or sensitive data can be unintentionally revealed.

Insecure plugin design
Poorly designed plugins can allow malicious requests or execution.

Excessive agency
The model performs unintended actions due to over-permissioned functionality.

Overreliance
Blind trust in AI outputs can lead to errors, security breaches, or legal risks

Insecure output handling
Poorly validated or sanitized model outputs can create security vulnerabilities.

So, what's the best defense for my AI?

You can safeguard your AI system with these steps:

ADVERSARIAL TRAINING

Equip your model to handle malicious or unexpected inputs.

- Train using adversarial data to improve the model's defenses.
- Use open-source tools designed for this purpose, including those for identifying and mitigating vulnerabilities.

MODEL EVALUATION

Test your model thoroughly before deployment.

- Simulate potential attacks to spot weaknesses (red team exercises).
- Stress-test in real-world and edge-case scenarios.
- Benchmark your model against known adversarial attacks to measure resilience.

INPUT VALIDATION AND SANITIZATION

Prevent harmful inputs from reaching your model.

- Validate and sanitize all inputs.
- Apply allowlists/blocklists to control what the model processes.
- Use tools like input fuzzing to test reactions to unusual inputs.

CONTENT MODERATION AND FILTERING

Ensure outputs are safe, ethical, and aligned with your goals.

- Use moderation tools to block inappropriate or harmful content.
- Regularly audit outputs for bias or compliance issues.
- Let users report problematic outputs to improve moderation policies.

DATA INTEGRITY AND PROVENANCE

Secure the trustworthiness of your training and input data.

- Verify data sources to prevent tampering or poisoning.
- Track data origins with provenance tools.
- Monitor live data flows for suspicious changes or unauthorized access.

ACCESS CONTROL AND AUTHENTICATION

Restrict access to your model and sensitive components.

- Assign user roles to limit permissions.
- Use multi-factor authentication for added security.
- Log and audit access attempts to detect anomalies.

SECURE MODEL DEPLOYMENT

Deploy your model with security measures in place.

- Isolate the model in a controlled environment.
- Regularly update the model and infrastructure to fix vulnerabilities.
- Monitor its behavior in production to detect anomalies.

AI GOVERNANCE

Ensure that AI models are developed and managed responsibly.

- Set clear policies for model use and monitoring.
- Maintain transparency in decision-making and data usage.
- Regularly update governance frameworks to align with evolving AI regulations.

And don't forget about AI regulations

From our experience, AI project requirements often depend on the region, industry, and legal standards. Knowing these upfront is important, as they can impose serious limits on a project's feasibility or features.

As for 2024, there are several **global jurisdictional considerations:**

USA

- Federal AI Disclosure Act requires AI-generated content to be clearly labeled.
- Federal Algorithmic Accountability Act mandates companies to assess and report the potential risks of automated decision systems.
- Blueprint for an AI Bill of Rights establishes guidelines to protect individuals from AI-related harms.

CANADA

- Artificial Intelligence & Data Act regulates the development and deployment of high-risk AI systems.

EUROPE

- AI Act sets rules for AI development and implementation in the EU, categorizing systems by risk and ensuring safety and rights protection.
- AI Liability Directive eases the burdens for individuals to claim compensation for damages caused by AI systems.
- Liability for Defective Products updates the product liability framework to cover damages.

And **general legal considerations** on AI security.

What's it all about?
Some of the regulations may not directly target AI, but they govern the data on which AI systems rely.

A necessity or an option?
The standards aren't legally binding, but they're widely respected.

What's at stake?
Complying with these regulations, enhances credibility, safeguards brand reputation, and builds trust with customers and stakeholders

PRIVACY GLOBAL/LOCAL LAWS:

- GDPR (EU) regulates data processing and protects user privacy rights.
- CCPA (USA) sets rules for data privacy and consumer rights in California.
- HIPAA (USA) governs the protection of health information in the healthcare sector.

PRIVACY STANDARDS

- ISO 31700 & ISO 29100 provide guidelines for privacy management and risk assessment in AI systems.
- ISO 27001 helps establish information controls within an organization's management systems.
- FIPS (USA) standards for securing federal information systems.
- NIST Privacy Framework (USA) offers guidelines to manage privacy risks and align with emerging regulatory frameworks.

OWASP SECURITY BEST PRACTICES:

- While not a legal requirement, OWASP offers guidelines and tools to help identify and mitigate AI security risks and strengthen system protection against vulnerabilities.

Now, let's turn all the knowledge we've got into a solid defense plan

In any major endeavor, planning is key — same with AI. Addressing security concerns upfront lays the groundwork for a secure and efficient development process.

At Aristek, we use the following checklist to ensure all critical aspects are covered **from the very start.**

Here's what we check:

Are there laws or regulations you must follow?
Certain AI applications, like facial recognition, are restricted by regulations such as the EU AI Act. We help ensure compliance from the get-go.

Can you ensure transparency?
Will you be able to clearly explain how the AI system makes its decisions? We focus on creating explainable AI that builds trust.

Is the data usage lawful?
Do you have permission to use the data, particularly if it includes personal information collected for another purpose? We guide you on privacy rights and permissions.

Can you safeguard privacy rights?
Our team ensures that robust privacy policies are in place, respecting rights like access, correction, and erasure of personal data.

Can unwanted bias be reduced?
We work to minimize biases in your AI systems, ensuring fairness and avoiding discriminatory outcomes.

Do you have the right expertise?
Do you have the required data science and AI skills within your team? We provide the necessary guidance and expertise to strengthen your project.

Can unwanted behaviors be controlled?
Is there a plan to avoid unintended or harmful actions? We'll analyze it.

Build your AI, secured

Then, we bring all this knowledge into action.
Security is always a top priority in AI solutions we build. Our AI systems don't just work — they protect, empower, and keep your data safe and sound.

Contact us →