

securing

Ai

key risks, threats, and countermeasures
for enterprise resilience

December 2024

**Booz
Allen® Ai™**

contents

introduction: critical systems, singular risks	1
why is Ai security unique—and why is it critical?.....	2
who is responsible for Ai security in the enterprise?	3
how Ai security engineers reinforce the security foundation	4
what do Ai security threats look like?	5
getting started with an Ai security strategy	8
everyone has a role to play in Ai security	10
appendix a: mobilizing to advance Ai security resources.....	12
about the authors	13
endnotes	13



introduction: critical systems, singular risks



With the stratospheric rise of large language models (LLMs), virtual assistants, predictive analytics, and more, organizations have entered a new era of computing in which artificial intelligence (AI) is becoming ubiquitous throughout the enterprise. As this powerful technology forever alters practices across industries, a new cyber threat landscape is also coming into focus. The accelerating adoption of often enigmatic AI systems introduces new threats and vulnerabilities that place enterprises at risk, motivating leaders in all sectors to revisit the security status quo.

Here's one reason why: AI systems are increasingly responsible for supporting strategic decision making. It's no surprise threat actors are building algorithmic and mathematical methods expressly designed to degrade, deny, deceive, or otherwise manipulate these mission-critical AI systems.

The more AI-driven use cases an organization implements—and the more broadly it scales automation—the more AI integrations fall into the crosshairs of adversaries, bringing increased risk of mission failure and reputational damage. Both the danger and complexity of these risks underscore why AI security is top of mind for chief technology officers (CTOs), chief information security officers (CISOs), and chief artificial intelligence officers (CAIOs) alike.

Not long ago, the rush to move infrastructure to the cloud increased cyber risks, and as the attack surface grew, enterprises fought to implement resilient endpoints. Today, in the wake of rapid AI adoption, the same commitment and agility are needed to protect the enterprise. However, organizations already face a shortage of AI and cybersecurity

talent, and these gaps are growing due to the blistering pace of innovation. Finding professionals well-versed in the intersection of these skillsets is like finding the proverbial needle in a haystack.

For example, the [2024 ISC2 Cybersecurity Workforce Study](#) surveyed more than 15,000 cybersecurity practitioners and decision makers. The respondents identified AI as their number-one security skill gap. Ben Aung, chief risk officer at the business-software firm Sage, [explained to the Wall Street Journal](#) that those who understand the security risks of LLMs and can collaborate effectively with data scientists and AI engineers “. . . is a much smaller and rarer group of people.” O'Reilly's [State of Security in 2024 study](#) found similar concerns, with 33.9% of respondents identifying a shortage of AI skills, particularly for vulnerabilities like prompt injection, as one of their most significant gaps.

By creating a centralized cadre of AI security experts and empowering them to collaborate with business and IT stakeholders, organizations can maximize the reach and effectiveness of this limited pool of expertise. This approach will allow them to develop the organizational structure, best practices, and expertise needed to stay ahead of these rapidly expanding and often invisible threats.

Additionally, with a cross-functional governance approach, an awareness of the latest threats, and the staffing and ability to implement calibrated countermeasures, organizations can confidently turn an AI-first posture into a secure mission transformation that builds new levels of long-term trust with stakeholders in the AI era.

why is **Ai security** unique— and why is it critical?

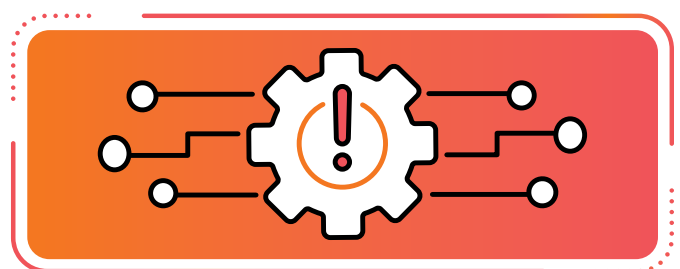
Many enterprises have made strides in enhancing their cybersecurity posture with strategies including defense-in-depth, secure-by-design, and zero trust. These efforts provide a solid foundation for securing AI systems. However, significant differences in how AI systems are built and exploited compared to conventional systems require additional new approaches.

Consider that the AI models being deployed today are the most complex software modules ever created—and that this intricacy can make finding, analyzing, and remediating attacks more difficult than with traditional software systems. As a unique technology, AI undoubtedly poses special security challenges. Here are just a few:

- **The AI “black box” hides critical cybersecurity information.** A complex machine operates inside a transparent box—but now the glass has been blacked out. After the switch is flipped, it’s often a mystery how inputs become outputs. Given the explainability challenges posed by generative AI (GenAI) models, the user can guess about what can’t be “seen,” but it isn’t feasible to dissect that hidden process. As models evolve from simple linear regressions to billion-parameter neural networks, it becomes that much more difficult to grasp through human analysis or intuition what the AI is doing. While this dynamic results in more powerful models that can be applied to more generalized problems, it puts safety and security professionals at a disadvantage who could benefit from visibility into how applications function and how these applications interact with networks, systems, and devices—not to mention how they use data.
- **Third-party, pretrained, and open-source AI models conceal sizeable risks.** Since most enterprises can’t develop AI models comparable to the leading models today, they procure them from third-party providers, believing robust cyber tooling comes built in. But that isn’t always true: Providers likely define security and risk differently than enterprises, focusing on systemic vulnerabilities and industry compliance, while enterprises prioritize mitigation of immediate risks tied to specific use cases. In this respect, it’s best practice for users to provide their own security layer. Similarly, an organization’s model engineering team may download a pretrained, open-source product. However, malware and other threats can be overlooked. Again, the organization will need to find its own security solution, potentially by working with an integrator.

- **The distributed and ubiquitous nature of AI magnifies its vulnerabilities.** The growing adoption of AI as autonomous assistants creates additional risks, as these AI agents are increasingly embedded into business processes with limited oversight. These threats and risks are amplified by the way AI’s decentralized acquisition and deployment posture removes its center of gravity. Furthermore, employees may be doing shadow AI—that is, using AI tools at work without the explicit approval and oversight of governance and IT teams—while vendors and other third parties are embedding AI into products without notifying users, all of which decreases transparency and weakens security guardrails. Malicious actors will seek to exploit the attack surface by probing overlooked entry points and interdependencies.
- **Anomaly detection loses much of its forensic usefulness.** When safeguarding an AI model, organizations can’t rely as much as they normally would on the detection of anomalous activity to discover and analyze malicious intent. Why? AI systems, by design, often act stochastically and produce unpredictable results. The non-deterministic nature of GenAI—that is, its “creative” ability to give different answers to the same prompt—adds complexity to security testing by multiplying error scenarios. With no decision tree, engineers can’t assess every imaginable attack path. While research into these emerging vulnerabilities has recently intensified, much remains unknown.

Given these and other AI-related challenges, most organizations will need to implement multiple defenses to protect their systems. An integrated set of strategies addresses how malicious actors target different phases in the AI system lifecycle with different attacks, how initial defenses often must evolve as AI itself is used to launch attacks, and how AI’s rapid evolution accelerates threat emergence. Critical to mounting this defense is cross-functional collaboration with shared security responsibility.



who is responsible for Ai security in the enterprise?

As AI becomes more integral to many enterprise systems, the challenge of securing these systems grows more complex. This raises an important question: How should the existing system security paradigm adapt to the risks AI introduces?

For example, a Gartner® report states “[n]early all (93%) of IT/security leaders surveyed are at least somewhat involved in their organization’s GenAI security/risk management efforts, but just 24% said they own this responsibility.”¹ Separately, a recent [IBM Institute for Business Value](#) (IBM IBV) survey found that while 82% of respondents say secure and trustworthy AI is essential to the success of their business, just 24% of their current generative AI projects have a component to secure these initiatives.

Despite their unique characteristics, AI systems remain enterprise systems, and they must adhere to established security requirements. Traditional risk, DevSecOps, and cybersecurity teams will continue to play a crucial role in securing AI systems.

However, they will be increasingly supported and augmented by AI security engineering teams with the expertise needed to peer more deeply into core AI operations and special risks. As just one example, identifying potential vulnerabilities requires understanding how the technology operates and the interdependencies created by specific use cases.

As a result, we are seeing that the responsibility for ensuring the security of AI systems within an enterprise is multifaceted, requiring both traditional security measures and specialized AI security expertise (Figure 1).

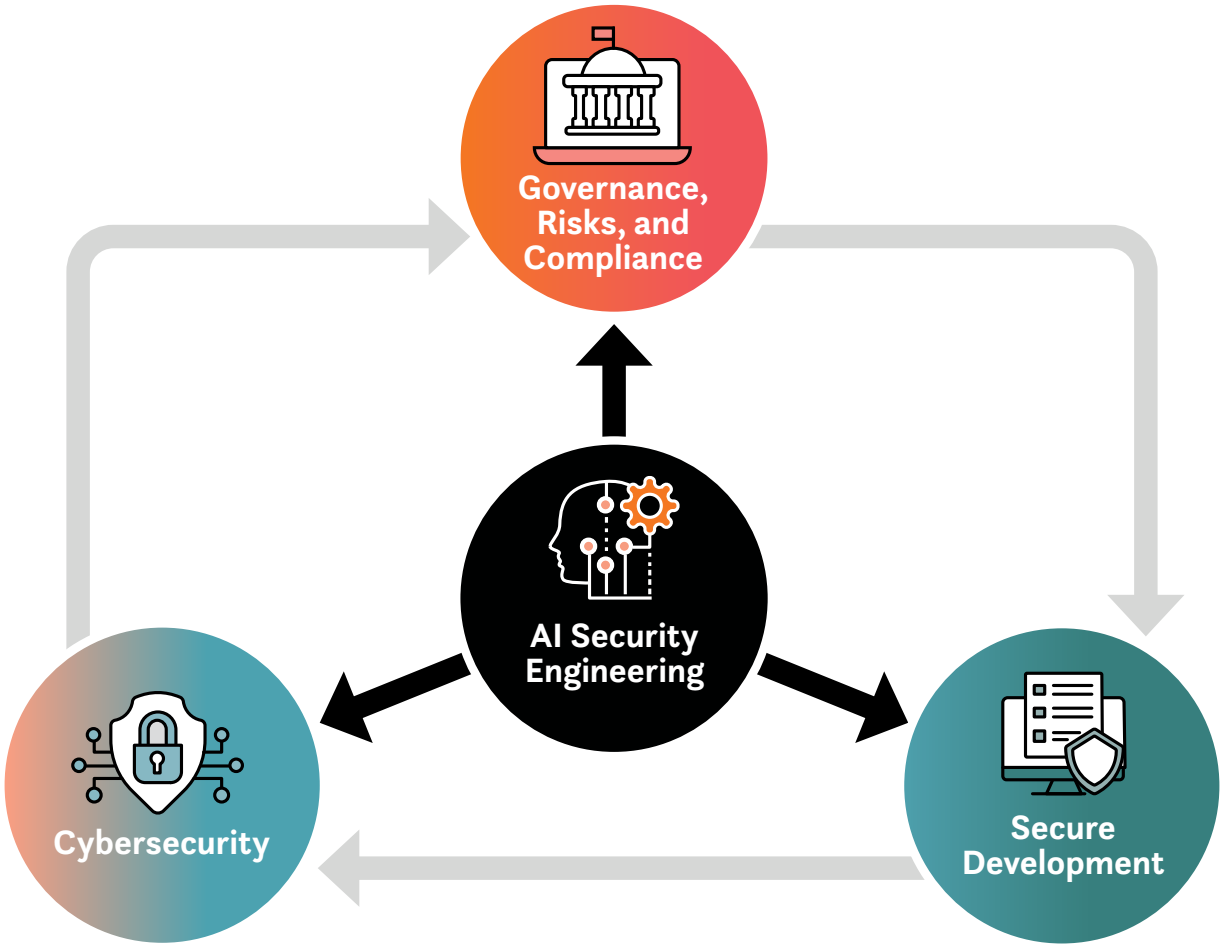


Figure 1: AI security engineering brings critical insight into AI operations across all facets of security operations

how **Ai security** engineers reinforce the security foundation

To build and sustain truly secure systems, leading enterprises employ an integrated approach that assesses potential risks, implements critical guardrails, hardens systems during design and development, and actively monitors and defends against potential threats. Spanning the entire system lifecycle, the primary responsibilities include:

- **Governance, Risks, and Compliance (GRC):** Ensuring continued trust and confidence in computing requires that organizations actively assess and mitigate potential risks and establish procedures to comply with regulatory requirements and organizational policies. While these roles may be distributed across various functions—for example, auditors employed by compliance and IT security—their actions should be integrated into a comprehensive GRC framework to maintain full coverage.

AI security engineers can enhance GRC for AI-related risks by monitoring for data privacy and other issues, assisting with Federal Information Security Management Act compliance, and collaborating to integrate AI risk management with broader GRC guardrails.

- **Secure Architecture and Development:** Secure-by-design ensures that security is an integral part of the system’s architecture, reducing vulnerabilities and potential attack surfaces. By prioritizing security from the outset, enterprises can build AI systems that are inherently more resilient and trustworthy. This proactive approach not only protects against known threats but also helps mitigate future risks by establishing a strong security foundation.

AI security engineers can ensure more robust model training and architectures to strengthen their attack resilience.

- **Cybersecurity:** Defense-in-depth is a fundamental cybersecurity strategy that applies multiple layers of defense throughout the system to protect against a variety of threats. For AI systems, this approach is particularly important due to the complex and evolving nature of AI-related risks. This multi-layered strategy includes continuous monitoring, regular security assessments, and adaptive defenses that evolve with emerging threats.

AI security engineers can perform deep testing and monitoring of AI systems to detect suspicious behavior that might be otherwise overlooked due to the technology’s non-deterministic nature while also addressing risks caused by LLM outputs being used as inputs to other systems.



To build and sustain truly secure systems, leading enterprises employ an integrated approach that assesses potential risks, implements critical guardrails, hardens systems during design and development, and actively monitors and defends against potential threats.

While traditional security measures provide a strong foundation, the unique challenges posed by AI systems necessitate specialized expertise. This is where the emerging discipline of AI security engineering comes into play. AI security engineering augments each phase of the security lifecycle with specialized knowledge and techniques tailored to AI systems.

what do **Ai security** threats look like?

AI security lapses and glitches take various forms, and even the world's most well-funded AI innovators are not immune. Reputational damage, and in some cases, financial liability have resulted. Examples include:

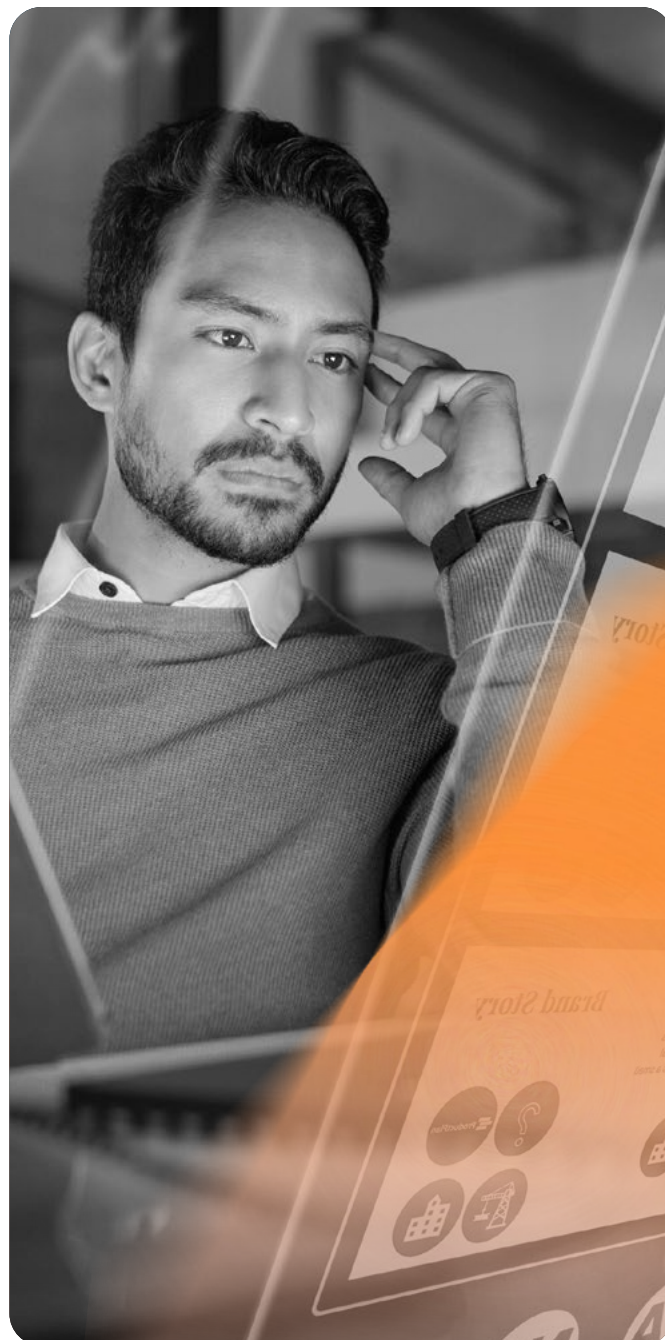
- The TrojanPuzzle attack forces code generation models to suggest insecure, vulnerable, and malicious code.
- ChatGPT leaks sensitive training data when it is asked to repeat the word “poem.”
- The LAION-5B dataset used to train Stable Diffusion is found to contain hundreds of illegal images.
- Pytorch-nightly is compromised when a malicious dependency package exposes sensitive information on Linux machines.
- Two individuals fool the Shanghai government's live facial authentication model to steal \$77 million from the tax system.

Representing a variety of attack vectors, these malicious actions and algorithmic errors highlight the critical need for organizations to implement controls that allow safe integration of AI with core systems and processes.

Understanding Common AI Security Threats


Who is attacking AI? Adversarial personas range from individual threat actors and hacktivists to financially motivated criminal organizations and nation-states that seek to undermine other nations' intelligence, military, and decision-making capabilities. And while some attacks have highly malicious intent, others simply represent attempts to game the system for personal gain.

While we have yet to experience a catastrophic attack on AI systems, many experts would counter that it's a matter of “when” not “if” it happens—given the value of the information underlying these systems and their overall risk profile. As just one factor, a comparatively smaller number of AI systems are deployed in full production, given challenges in certifying them as fully secure to achieve authority to operate (ATO). While these limited deployments might provide short-term protection, it is not a sustainable operating model, as enterprises fail to benefit from AI in the interim. Rather, building appropriate protection is fundamental to using AI.



Although AI model attacks may not always produce the magnified impacts threat actors seek, the attack surface continues to be inviting, as these types of attacks may in some cases be easier to carry out than breaking into web servers or overwhelming networks.

Here are five attacks against AI systems that can significantly harm the enterprise:


 Data Poisoning	Adversaries manipulate training data to compromise model behaviors and insert backdoors.		
	Attack Occurs During Model Training	Attack Occurs Between Model Training and Model Inference	Attack Occurs During Model Inference
	●		

Models are only as effective and useful as the data they're given, and models trained on poisoned data—data that adversaries have tampered with to intentionally manipulate model performance—can behave in unexpected ways. That's why security-conscious organizations often view data poisoning as a top security concern, and why adversaries see it as an attractive attack mode.

Data poisoning encompasses multiple types of manipulations. Adversaries may covertly add noise to training datasets, change labels, or add entirely falsified or cherry-picked data to skew decision boundaries favorably for them. In the

financial sector, for example, data poisoning poses a significant threat to algorithmic trading models. Adversaries can inject deceptive data into the training set to compromise the model's predictive accuracy, potentially leading to financial losses from falsified price forecasts. Alternatively, a regulatory agency may worry about data poisoning by omission, where inside threat actors are motivated to remove the information machine learning models need to flag evidence of potential fraud.

To address data poisoning, organizations should work to ensure the integrity of training sets.


 Malware	Adversaries package malicious code within model files and libraries.		
	Attack Occurs During Model Training	Attack Occurs Between Model Training and Model Inference	Attack Occurs During Model Inference
		●	

Malware in an AI model is like malware in any other file—malicious code that cybercriminals insert with intent to steal information, destroy system functioning, or facilitate ransomware attacks.

An organization that downloads a pretrained LLM from the internet—possibly as the foundation for a fine-tuned model—may discover malware embedded in file formats,

some of which are inherently vulnerable. As vulnerabilities emerge (e.g., Fog ransomware), the industry adopts precautions, but threat actors also innovate toward new forms of malware.

Organizations can address this threat by scanning these files with model-specific virus scanners, as traditional scanners are not designed for these formats.

 Model Evasion	Adversaries perturb model inputs to control model outputs.		
	Attack Occurs During Model Training	Attack Occurs Between Model Training and Model Inference	Attack Occurs During Model Inference
			●

In simple terms, model evasion means engineering an input to fool the model and control the output or otherwise force the model to behave in unintended ways. An example might be placing a small sticker on a stop sign—an almost imperceptible perturbation of the sensory environment—to trick an autonomous vehicle into driving forward. In fact, it is remarkably easy to fool

image classification models. It's not just that adversaries can cause a random misclassification—they can ensure that the model sees what they want it to see.

To address model evasion, engineers can train models to be robust to perturbed inputs or train them to recognize, classify, and handle such inputs.

Data Leakage and Model Theft



Adversaries infer and steal sensitive training data, model behavior, and/or intellectual property.

Attack Occurs During Model Training

Attack Occurs Between Model Training and Model Inference

Attack Occurs During Model Inference



Privacy attacks aim to steal model information. Data leakage, also known as “model inversion,” is stealing the training data, while model theft is stealing the model itself (often to avoid the costs associated with data collection and training and to craft better evasion attacks against the target model).

Data theft has the potential to be highly damaging given the sensitive nature and massive amount of the data that tech companies, governments, healthcare organizations, financial services firms, and other enterprises gather and use to train their AI models. This damage can include

privacy breaches, financial loss, compromised national security, and service disruptions. Sometimes LLMs share information they weren’t supposed to reveal, including their own training data. This training data can be more valuable to threat actors than confidential information, since it can provide clues and insights to enable model evasion attacks.

To address some data-related threats, organizations can implement differential privacy measures, which provide a statistical framework that anonymizes an individual’s data while still allowing for information sharing and analysis.

Large Language Model Misuse



Adversaries override an LLM’s instructions and safety alignment.

Attack Occurs During Model Training

Attack Occurs Between Model Training and Model Inference

Attack Occurs During Model Inference



LLM-powered AI chatbots are increasingly being integrated into enterprise systems as both a guide and a user interface. What are the security implications of this? One concern is that the LLM, having been trained on the internet at large, will somehow output information it shouldn’t, such as toxic or dangerous content (e.g., instructions to build a bomb). To minimize this type of harm, model providers engineer safety controls into their systems. Broadly speaking, jailbreaking is adversaries’ intentional use of specific techniques to neutralize or bypass these safety controls to elicit this harmful content and expose the LLM to further manipulations.

Heightening the risk is the transferability property of machine learning models—or the way an attack demonstrated to work against one model will work against a similar model. Skeleton

Key, for example, is a sophisticated cyberattack where hackers use a multi-step process to circumvent the security protocols of AI systems, enabling them to generate harmful or unauthorized outputs. With its robust transferability, this attack damages multiple GenAI models, across OpenAI, Anthropic, and Llama.

To reduce the risk of jailbreaking and prompt injection, organizations can use input validation to filter harmful prompts, limit the number of queries a user can make, and monitor query sequences to block suspicious patterns. Alternatively, to address this threat without limiting queries, organizations can rigorously monitor and assess outputs to prevent the LLM from making errors with what information it reveals.

It is remarkably easy to fool image classification models. It’s not just that adversaries can cause a random misclassification—they can ensure that the model sees what they want it to see.

getting started with an Ai security strategy

Given the diversity of threats and threat actors, enterprises cannot simply embrace a one-size-fits-all strategy. Rather, a tailored, risk-based approach is needed to mitigate their unique risks effectively and safeguard their proprietary AI systems.

In working with enterprises large and small, Booz Allen often uses a comprehensive security framework such as

MITRE ATLAS to benchmark the current state and identify realistic objectives that meaningfully enhance the security posture. A next step is to gain a broad understanding of the many available best practices for AI security engineering that reduce the likelihood of harmful model outcomes.

The strategies, approaches, and operating models that we have developed or adapted for AI security include:

 <p>Planning</p>	<ul style="list-style-type: none"> • Risk Modeling: Identify risks and quantify the probability and impact of a risk realization.
 <p>Measurement</p>	<ul style="list-style-type: none"> • Red Teaming: Uncover vulnerabilities, weaknesses, or threats by simulating realistic attack scenarios. • Security Testing: Quantify the likelihood that a model’s training data was poisoned and/or quantify model robustness to adversarial attacks (e.g., data omission detection, data leakage measurement).
 <p>Security Engineering</p>	<ul style="list-style-type: none"> • Model Scanning: Prevent the introduction of malicious code by inspecting model files for unexpected patterns. • Dependency Scanning: Scan model training source code for known vulnerabilities in libraries and dependencies. • Data Tampering Detection: Establish checks across the machine learning operations (MLOps) lifecycle to ensure only “original” data is for model training and evaluation (e.g., data integrity checks via cryptographic signatures). • Robust Model Training: Ensure that models protect sensitive data, are resilient against adversarial attacks, and maintain data confidentiality (e.g., differentially private training).
 <p>Operations</p>	<ul style="list-style-type: none"> • Operational Controls and Monitoring: Enforce model security during operations, typically by implementing controls that monitor or manipulate model inputs and/or outputs (e.g., prompt injection or jailbreaking detection). • Model Updates: Patch discovered vulnerabilities, refine risk assessments, and assess the need to perform comprehensive model audits.
 <p>Control</p>	<ul style="list-style-type: none"> • Governance: Establish procedures, guidelines, best practices, authorities, and mitigations that align to the organization’s MLOps lifecycle.

Figure 2: Specific tactics aligned across the AI security lifecycle

In assessing these options, organizations should conduct a tailored analysis to identify what to incorporate rather than setting out to apply each practice. This partly means assessing unique enterprise drivers. Specifically, organizations should think through the specific cost pressures, time-to-market demands, and requirements for stakeholder buy-in that will shape the strategy.

At a minimum, organizations must qualify and control risks by performing risk modeling and establishing a governance plan (Figure 3). Most should also plan to carry out red teaming and security testing protocols that allow risks to be quantified. In addition, open-source tools for model and dependency can be incorporated at very low cost and quickly integrated into the MLOps pipeline. Additional components of the AI security engineering strategy will likely depend on the model’s maturity level and use-case type. As a baseline, these risk profiles can be assessed as follows:

- 1. GenAI for Business:** Standalone, commercial GenAI offerings, such as Microsoft Copilot, have the potential to improve business productivity, but they also introduce vulnerability to misuse and unexpected output. While risks are lower than with customized models, these systems often have access to sensitive systems and data. As such, governance protocols should outline the appropriate use and monitoring of GenAI systems.
- 2. Third-Party GenAI Model Integration:** Third-party GenAI model integration is becoming increasingly common for organizations interested in embedding LLMs into their applications. Because of the sensitivity of these interactions, organizations should thoroughly examine

third-party models for vulnerabilities and quantify risk. This process may include risk modeling and red teaming, and integration should be guided by governance protocols and subject to monitoring.

- 3. Pretrained Models:** Models built and trained by outside organizations can be powerful tools, but they may not come with in-built AI security measures. Organizations should comprehensively test these models before implementation; this process can include model and dependency scanning, operational controls, and red teaming.
- 4. Fine-Tuned Models:** Externally developed models fine-tuned in house will have a similar risk profile to pretrained models. They should be subjected to the same protocols and scrutiny. In addition, organizations should examine the data used to fine-tune these models for tampering—especially if it is scraped from the internet.
- 5. Homegrown Models:** Homegrown models can have varied risks depending on use case and architecture, so it’s important to integrate AI security into the design process from the beginning. This can include risk modeling, governance protocols, robust model training, and red teaming.

With a customized, risk-based approach and the right integration of individual strategy components, organizations can expect to dramatically strengthen their AI security posture and threat mitigation capacity in a relatively short time, even as adversaries continue to deploy shifting tactics and techniques.

	Practice	GenAI for Business	3rd-Party (GenAI) Model Integration	Pretrained Model	Fine-Tuned Model	Homegrown Model
Planning	Risk Modeling	MUST	MUST	MUST	MUST	MUST
Measurement	Red Teaming	WON'T	SHOULD	SHOULD	SHOULD	SHOULD
	Security Testing	WON'T	WON'T	WON'T	COULD	SHOULD
Security Engineering	Model Scanning	WON'T	WON'T	MUST	MUST	MUST
	Dependency Scanning	WON'T	WON'T	WON'T	MUST	MUST
	Data Tampering Detection	WON'T	WON'T	COULD	COULD	COULD
	Robust Model Training	WON'T	WON'T	COULD	COULD	COULD
Operations	Monitoring	COULD	SHOULD	SHOULD	SHOULD	SHOULD
	Detection	COULD	SHOULD	COULD	COULD	COULD
	Manipulation	WON'T	WON'T	COULD	COULD	COULD
Control	Governance	MUST	MUST	MUST	MUST	MUST

Figure 3: Using the MoSCoW method to identify *Must*, *Should*, *Could*, and *Won't* security requirements based on a hypothetical organization’s unique risk profile—this graphic is a representative mapping that should not be viewed as a universal guide.

everyone has a role to play in **Ai security**

Ultimately, AI security must be a team effort. CIOs must ensure appropriate policies are in place, chief risk officers need to maintain vigilant oversight, CTOs should instill disciplined development approaches, and CISOs must protect operating environments using a zero-trust mindset. However, it also increasingly requires that AI security engineering expertise be infused across each of these areas.

The inclusion of AI technologies within enterprise systems requires that organizations update some of their most critical operating policies, including:

- **Integrated Oversight and Management:** Clear communication channels and defined responsibilities, including established protocols for sharing new risk information, are essential to ensure that all aspects of AI security are addressed comprehensively.
- **Risk Assessment:** Identifying potential risks requires a detailed understanding of specific AI use cases, as the nature of these risks varies dramatically from traditional technologies.
- **Governance and Policy:** New policies and requirements are needed to address specific AI-related risks, including monitoring and enforcement.
- **Continuous Improvement:** Best practices for lifecycle security include continuous monitoring, regular security assessments, and regular model updates—a role that AI security engineering is well-positioned to lead.
- **Upskilling:** The dynamic nature of AI technology necessitates continuous learning and improvement in security practices. Investing in training for security teams, staying informed about emerging threats, and incorporating new security measures are vital for maintaining robust AI security.

By integrating secure-by-design principles, continuous monitoring, and cross-functional collaboration, enterprises can ensure the security and trustworthiness of their AI systems. As AI technology continues to evolve, so must the strategies and practices for securing it, ensuring that AI systems remain resilient against both traditional and AI-specific threats.

Booz Allen stands ready to help you navigate these challenges with confidence, stay ahead of emerging threats, and realize the operational and mission benefits of AI systems while mitigating risks.





appendix a: mobilizing to advance

Ai security resources

Amid the bad news about AI’s vulnerabilities, there is also reason for optimism: The expansion of AI security threats has catalyzed a corresponding growth in tools for countering them. As CISOs and other security leaders work to qualify, quantify, and collaborate with business counterparts to mitigate AI risks, they now have access to a vibrant stakeholder ecosystem with more resources available than ever before.

Together, the new frameworks, tools, and insights testify to a recent burst of collective energy that has significantly strengthened organizations’ ability to operationalize, scale, and harmonize effective AI security protocols. Booz Allen can help organizations distill these approaches and insights into tailored real-world strategies. Examples of helpful resources include:

Regulators	<ul style="list-style-type: none">• The Biden Administration’s Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence calls for robust, reliable, repeatable, and standardized evaluations of AI systems before they are operationalized.
Analyst and Research Firms	<ul style="list-style-type: none">• Gartner’s AI TRISM Framework provides a structured approach to help organizations assess and manage the risks associated with the deployment of AI systems.
Standards Organizations	<ul style="list-style-type: none">• The National Institute of Standards and Technology’s (NIST’s) Trustworthy and Responsible AI Publication (NIST AI 100-2e2023) defines a taxonomy and terminology for adversarial machine learning attacks and mitigations.• NIST’s Artificial Intelligence Risk Management Framework (NIST AI 100-1) offers a path to minimize potential negative impacts of AI systems.• NIST’s Artificial Intelligence Safety Institute Consortium (AISIC) helps ready the nation to address the capabilities of the next generation of AI models or systems with appropriate risk management strategies.
The Safety and Security Communities	<ul style="list-style-type: none">• MITRE ATLAS is a knowledge base of adversary tactics and techniques.• The Open Worldwide Application Security Project (OWASP) Top 10 for LLM Applications project aims to elucidate potential security risks when deploying and managing LLMs.• A tool called huntr is a community-supported, public repository for identifying AI/machine learning application vulnerabilities.• The AI Vulnerability Database (AVID) is an open-source knowledge base of failure modes for AI models, datasets, and systems.• MLCommons provides a host of tools for measuring and improving the accuracy, safety, speed, and efficiency of AI technologies.
Industry	<ul style="list-style-type: none">• Google’s Secure AI Framework (SAIF) provides methodologies for integrating ML model risk management, security, and privacy across AI systems.

about the authors

- **Justin Neroda** is a senior vice president in Booz Allen’s AI business, leading the firm’s work with national security clients.
- **Matt Keating** leads Booz Allen’s Secure AI practice, a component of the firm’s AI business focused on adversarial AI and AI security solutions.
- **Dr. Andre Nguyen, Ph.D.**, is an adversarial machine learning (ML) expert within Booz Allen’s Secure AI practice, leading advanced research on threats and vulnerability within enterprise AI systems.
- **Shafi Rubbani** is a Booz Allen researcher and scientist with expertise in deep fake detection and hardening systems against data noise, model poisoning, and other attack vectors.

endnotes

- ¹ Gartner, [Generative AI Security and Risk Management](#), 3 August 2023. [GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.]

Booz Allen AiTM

about Booz Allen

Trusted to transform missions with the power of tomorrow's technologies, Booz Allen Hamilton advances the nation's most critical civil, defense, and national security priorities. We lead, invest, and invent where it's needed most—at the forefront of complex missions, using innovation to define the future. We combine our in-depth expertise in AI and cybersecurity with leading-edge technology and engineering practices to deliver impactful solutions. Combining 110 years of strategic consulting expertise with the perspectives of diverse talent, we ensure results by integrating technology with an enduring focus on our clients. We're first to the future—moving missions forward to realize our purpose: **Empower People to Change the World[®]**.