AI and cybersecurity: The new frontier of business resilience

February 2026

The better the question.
The better the answer.
The better the world works.

Shape the future with confidence

Contents

Executive summary

Artificial intelligence (AI) has transitioned from a novel capability to a strategic core of business and national security. Cybersecurity, once viewed as an operational safeguard, now underpins digital trust and enterprise resilience. As generative AI (Gen AI) democratizes advanced capabilities, security teams face a double-edged sword. AI amplifies defense through faster detection and response but simultaneously lowers the cost and complexity of attacks. Research from a leading IT service provider shows that organizations using extensive security AI and automation saved roughly US$1.9 million in breach costs compared to those that did not. A global OEM study in 2024 found that Security Copilot users achieved about a 30% reduction in mean time to resolve (MTTR) incidents , reinforcing the role of AI in modern cybersecurity. The benefits are therefore real and measurable. However, adopting AI safely requires governance, workforce upskilling and careful consideration of adversarial risk.

The report 'AI and cybersecurity: The new frontier of business resilience' examines AI's evolving role in cybersecurity, starting with the global landscape of models, providers and computer infrastructure. It analyzes the economics of AI adoption and the unprecedented capital inflows into AI ready data centers. Subsequent sections assess the top AI use cases across security functions, the emerging threats introduced by generative and autonomous AI, and a holistic framework for securing AI models, data and infrastructure. The report discusses regulatory developments such as the EU Artificial Intelligence Act and NIST's AI Risk Management Framework and concludes with strategic recommendations for chief information security officers (CISOs). Embedded throughout the report are concrete examples and case studies, such as the US$25 million deepfake scam that exploited AI generated video to defraud an engineering firm, illustrating how attackers weaponize AI. The report emphasizes actionable guidance: security leaders should move beyond awareness to build cyber decision intelligence, which combines real time visibility, scalable analytics and data sovereignty to enable informed decisions at machine speed.

CHAPTER 1

Global AI landscape

The computation race foundation models and leading providers

AI innovation today is dominated by foundation models large neural networks pre-trained on vast datasets and then adapted for diverse downstream tasks. Transformer based large language models (LLMs) developed by leading global technology providers remain the most visible examples. Two flagship multimodal models released in 2025 demonstrate advanced capabilities across text, images, video and code, alongside notable improvements in reasoning and safety. At the same time, an expanding ecosystem of open-source model families continues to evolve, while a new wave of startups is releasing smaller, application specific models optimized for targeted use cases. Hardware advances are a key enabler of this progress, with next-generation accelerator chips designed specifically for AI training and deployment, supported by large-scale cloud infrastructure providers.

Country level strategies and investment

The compute race is no longer confined to companies; nations are building sovereign AI infrastructure to secure economic advantage. According to TRG Datacenters' 2025 analysis of global AI data center capacity, the US leads with 187 AI clusters, the equivalent of 39.7 million H100 chips and 19.8 GW of power capacity making it the world's most AI-dominant country. China, despite having 230 clusters, lags in compute density with only 400,000 H100 equivalents and 289 megawatts (MW) of power capacity, reflecting export control headwinds. The United Arab Emirates and Saudi Arabia have invested heavily in AI chips, each operating fewer than 10 clusters but hosting over 23 million and 7.2 million H100 equivalent chips, respectively. France has amassed 989,000 AI chips and 2 GW of capacity, while India hosts eight clusters with 1.2 million H100 equivalents. Overall, global AI infrastructure investment reached US$200 billion in 2025.

Geopolitical competition underlies these figures. The US is accelerating capacity through mega-investments. Stargate's US$500 billion plan for 10 GW and other major deals. Europe and China are also scaling, with France, the EU and Beijing announcing multibillion-dollar AI funding.

Global AI strategies and investment

National strategies guided these AI investments. The US emphasizes domestic AI capabilities through initiatives like the American Truly Open Models (AToM) Project and Stargate, while also dominating in Graphics Processing Unit (GPU) design through leading chip-making firms. China's New Generation AI Development Plan targets a core AI industry worth RMB 300 billion by 2025 and positions the country to be the global AI innovation leader by 2030.

The European Union focuses on ethical and trustworthy AI, with an InvestAI plan to mobilize €200 billion India is also expanding its compute infrastructure. In 2024, the Indian government announced the IndiaAI mission to build sovereign compute clusters and is collaborating

with hyperscalers to deploy high performance AI hubs, reinforcing AI-driven cyber risk management. Middle Eastern nations like Saudi Arabia and the UAE are using petrodollar wealth to build AI ready data centers and attract global talent.

Economics of AI in cybersecurity

Quantifying AI's economic value in security is difficult because its benefits come from preventing losses and improving productivity, rather than generating direct revenue. Gartner estimates that by 2027, over 40% of AI-related data breaches will stem from the improper use of GenAI across borders, underscoring growing cybersecurity risks in AI systems and the need for stronger AI risk management. According to the Cost of a Data Breach Report (2025) by IBM , the average breach costs US$4.4 million, but organizations with extensive AI and automation capabilities save roughly US$1.9 million due to faster detection and reduced remediation effort. These savings reflect faster detection and reduced post incident remediation. Microsoft's Security Copilot trial demonstrated about a 30% reduction in MTTR and concluded that analysts spend an average of 2.7 hours per day on incident response, representing US$3.3 billion in labor costs across US security operations center (SOC) teams. Reducing false positives and automating incident triage frees analysts to focus on high value tasks and reduces burnout. Traditional ROI metrics (cost vs. savings) inadequately capture these benefits, so new metrics are emerging that avoid breach losses, measure the percentage of automated triage, reductions in false positives and improvements in analyst productivity. For example, a global OEM provider reports that AI driven security orchestration and automation (SOAR) can reduce alert volumes by 60% and incident response times by 50%, enabling analysts to handle more complex cases. These efficiency gains translate into tangible cost avoidance and improved resilience against cyber threats to AI systems.

The financial stakes extend beyond direct security ROI. The AI boom has driven unprecedented data-center investment and intensified cyberattacks, reshaping the global digital landscape.

For instance, the data center equipment and infrastructure spending reached US$290 billion in 2024 and is projected to exceed US$1 trillion annually by 2030. Major hyperscalers accounted for nearly US$200 billion of this capital expenditure (capex). These investments create a cascade through the value chain, benefiting chipmakers, contract manufacturers, power grid suppliers and heavy equipment companies.

The economics of GPUs and supply chains

GPUs have become a scarce resource powering AI. A leading chipmaker's dominance in high end accelerators has created geopolitical tensions; US export restrictions limit shipments of top tier chips (H100/H200) to China, forcing Chinese firms to develop domestic alternatives. The 2025 AI compute race has thus become a story of supply chains. US companies are signing long term contracts to secure future GPU capacity, while startups experiment with AI optimized ASICs and energy efficient architectures to reduce reliance on scarce chips. Countries like France and India are investing in manufacturing to gain autonomy. Meanwhile, efficiency innovations, low rank adaptation (LoRA), retrieval augmented generation (RAG) and federated learning would help reduce compute needs and address security challenges in AI systems but may paradoxically increase overall demand as more organizations adopt AI.

Governance and regulatory landscape

Regulators worldwide are increasingly recognizing the need to establish frameworks that govern the use of AI in various sectors, including cybersecurity. These frameworks aim to facilitate the ethical, transparent deployment of AI technologies in compliance with existing laws.

Key regulatory initiatives include:

India AI governance guidelines

The India AI governance guidelines provide a robust framework for the ethical deployment of AI technologies within the cybersecurity domain. These guidelines emphasize several key principles:

- **Ethical AI deployment:** Organizations are encouraged to adopt AI systems that are designed to be fair, transparent and accountable. This involves implementing algorithms that are free from bias and enabling AI decisions to be audited and explained.
- **Data privacy and security:** The guidelines mandate strict adherence to data protection laws, requiring that personal and sensitive data is handled in compliance with the Digital Personal Data Protection Act (DPDPA). Organizations should implement data minimization practices, collecting only necessary data and processing it.
- **Accountability mechanisms:** Organizations are required to establish clear lines of accountability for AI-driven decisions. This includes defining roles and responsibilities for AI governance, with designated personnel overseeing AI systems and their compliance with regulatory standards.
- **Stakeholder engagement:** The guidelines advocate for the involvement of diverse stakeholders, including industry professionals, legal advisors and civil society, in the development and deployment of AI systems. This collaborative approach helps bring multiple perspectives into consideration, enhancing the ethical deployment of AI.

NIST AI Risk Management Framework (AI RMF)

The National Institute of Standards and Technology (NIST) of the US released its AI RMF 1.0 in 2023 to help organizations design, develop and deploy trustworthy AI systems, strengthening enterprise-wide AI risk management. It emphasizes characteristics such as validity, reliability, robustness, security, privacy, transparency and fairness. The framework guides risk management across the AI lifecycle and underscores that AI systems have socio technical dimensions. The AI RMF aligns with existing frameworks like the Cybersecurity Framework and Privacy Framework, providing a common language for industry and regulators. Organizations are encouraged to implement continuous monitoring and assessment of AI systems to facilitate compliance with established standards.

EU AI Act and global regulations

The European Union's AI Act (formally adopted in 2024 and entering full force by 2026) is the world's first comprehensive AI law. It classifies AI systems by risk, imposing strict requirements on high risk systems, including robustness against tampering, incident reporting and documentation of security measures. Article 15 requires appropriate levels of accuracy, robustness and cybersecurity to guard against data poisoning and model tampering. The Act interacts with other European laws, such as the General Data Protection Regulation (GDPR) and the NIS2 Directive. The GDPR enforces strict data protection measures, including the right to explain automated decisions made by AI systems, while the NIS2 Directive enforces privacy, cybersecurity and data residency obligations. Also, the EU's InvestAI program promotes trustworthy AI.

China's AI regulation follows a proactive, top-down model, combining broad national policies with detailed, application-specific rules governing generative AI (GenAI), deep synthesis and algorithmic recommendation systems. The country's interim measures for the management of GenAI services, which have been fully applicable since their adoptions in 2022 and 2023, require providers to register their models, comply with content governance and verify generated content is accurately labeled and can be watermarked to prevent confusion.

Regulations and Guidelines

The United States

NIST AI RMF

- Trustworthy AI principles
- End-to-end risk management
- Socio-technical focus
- Aligned with cyber and privacy frameworks
- Continuous AI monitoring

European Union

EU AI Act

- Risk-tiered AI regulation
- Security, robustness and incident controls
- GDPR and NIS2 integration
- €200b InvestAI for trustworthy AI

India

AI Governance Guidelines

- Fair and transparent AI adoption
- DPDPA-aligned data protection
- Clear AI accountability roles
- Multi-stakeholder oversight

China

AI Regulatory System

- Top-down AI governance
- Generative AI registration rules
- Mandatory labeling and watermarking
- Algorithm recommendation oversight

Data sovereignty laws

Data sovereignty refers to the principle that data collected or stored in a country is subject to that country's laws and regulations. Organizations should comply with each jurisdiction's rules on data access, storage, processing and movement. Some laws restrict cross border transfers altogether, while others require retaining a local copy or demonstrating legal necessity for transfer.

Data residency (the physical location of data) and data localization (mandatory storage within borders) are subsets of sovereignty. Managing compliance may require distributed infrastructure, sovereign clouds or multi cloud strategies that help data stay within jurisdictional boundaries. Regulatory regimes such as Europe's GDPR, China's Cross Border Data Flow provisions and India's Digital Personal Data Protection Act (DPDPA) all impose specific requirements. The EU AI Act also sets data governance requirements for training datasets and technical solutions addressing AI specific vulnerabilities.

Various jurisdictions are implementing data sovereignty regulations that require organizations to store and process data within specific geographic boundaries. This necessitates the establishment of localized data centers and compliance with local laws, impacting how AI systems are designed and deployed to support long-term cyber resilience in the age of AI.

Industry standards and OWASP guidelines

OWASP (the foundation that works to improve the security of software through its community-led open-source software projects) Machine Learning Security Top 10 catalogues common vulnerabilities in AI systems, including input manipulation, data poisoning, model theft, supply chain attacks and insecure model deployment. OWASP's GenAI Security Project expands this to generative models, highlighting prompt injection, model hallucination, excessive agent autonomy and malicious training data. These guidelines provide concrete mitigation strategies (input filtering, context isolation, rate limiting and adversarial testing) that align with technical best practices.

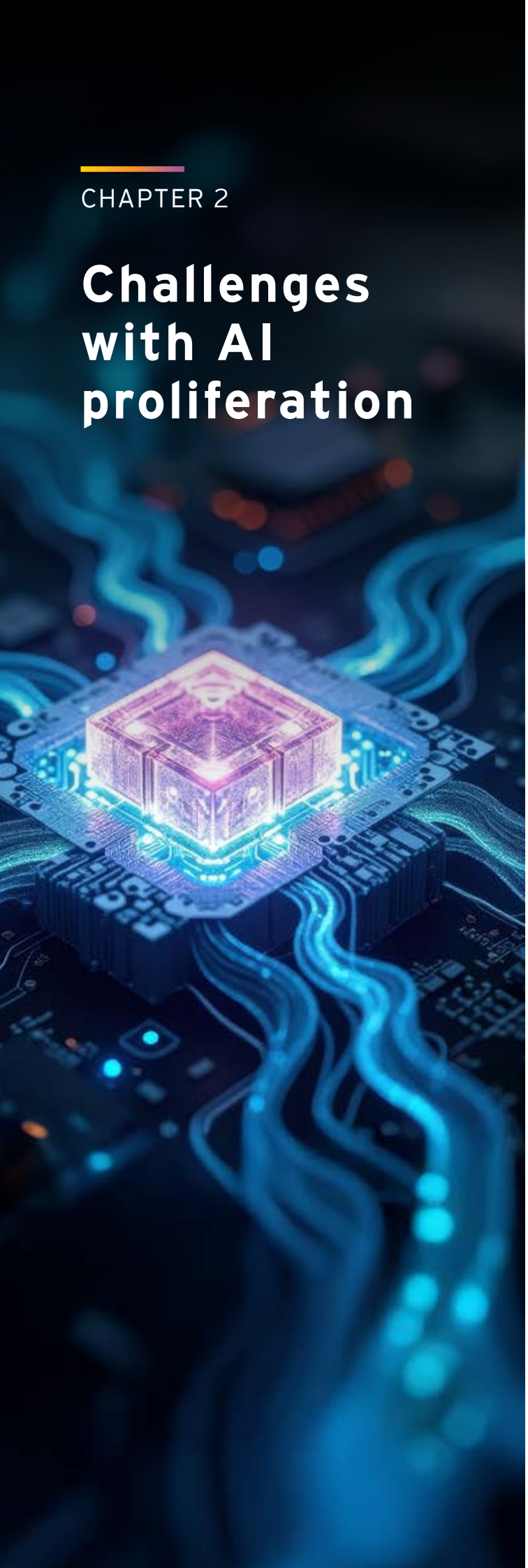
In addition to overarching AI regulations, specific industries are subject to their own regulatory controls. For example, the healthcare sector should comply with the Health Insurance Portability and Accountability Act (HIPAA) in the US, which governs the use of AI in managing patient data and enhancing privacy.

By strengthening these regulatory frameworks and implementing specific controls, regulators aim to foster trust in AI technologies while mitigating risks associated with their deployment. Organizations should stay informed about these developments and adapt their AI strategies accordingly to facilitate compliance and ethical use of AI in cybersecurity and beyond.

By incorporating these regulatory examples, organizations can better understand the landscape and improve compliance with evolving standards.

CHAPTER 2

Challenges with AI proliferation



As organizations increasingly adopt AI technologies in cybersecurity, several challenges have emerged that need to be addressed proactively. These include:

- **Lack of AI governance and visibility:** Many organizations struggle with establishing robust governance frameworks for AI deployment. This lack of oversight can lead to inconsistent practices and increased risks associated with AI usage.
- **Excessive token usage:** The absence of rate limiting and controls has resulted in multiple cases of excessive token usage, leading to increased operational costs and potential service disruptions.
- **Prompt hopping:** With numerous off-the-shelf products developing their own AI agents, there is a risk of prompt hopping, where users switch between different AI models to achieve desired outcomes, complicating governance and consistency.
- **Data exposure:** The integration of AI systems often involves handling sensitive data, raising concerns about data exposure and compliance with data protection regulations.
- **Response validation issues:** Models trained on synthetic data rather than real-world data may lead to response validation issues, where the AI's outputs do not align with actual operational scenarios.
- **Challenges with model protection:** Protecting AI models from adversarial attacks and intellectual property theft is increasingly difficult as AI systems become more complex and integrated into organizational workflows.

The transition for cyber paradigm

As we transition from the strategic role of AI in cybersecurity to the global landscape of AI models and computation, it is crucial to recognize the intricate relationship between these elements and the overarching theme of securing AI itself. The rapid evolution of AI technologies, particularly in the landscape of cybersecurity, underscores the necessity for robust frameworks that not only enhance defensive capabilities but also mitigate the risks associated with adversarial exploitation. The earlier chapter delved into the competitive dynamics of AI infrastructure, highlighting how nations and corporations are racing to secure their computational resources while simultaneously addressing the vulnerabilities that arise from deploying advanced AI systems.

This interplay between innovation and security is important; as organizations increasingly rely on AI threat detection and response, they should also prioritize the safeguarding of their AI models and data against emerging threats. By examining the foundational models, the race for computational supremacy and the geopolitical implications of AI investments, this analysis outlines a comprehensive approach to securing AI technologies, so that the benefits of AI in cybersecurity are realized without compromising safety and integrity.

Categorization of AI use cases

To effectively leverage AI in cybersecurity, organizations should first assess their strategic objectives concerning AI capabilities. This assessment should lead to the categorization of organizations into two primary groups:

- **Consumers of AI use cases:** These organizations utilize existing AI solutions developed by third-party vendors. They focus on integrating these solutions into their cybersecurity frameworks to enhance threat detection, incident response and overall security posture. Examples include deploying AI-driven Security Information and Event Management (SIEM) systems and endpoint protection solutions.
- **Leveraging AI to build enterprise-specific use cases:** Organizations in this category actively develop proprietary AI applications tailored to their unique cybersecurity challenges. This may involve creating custom machine learning models for threat detection, developing predictive analytics for vulnerability management or building automated incident response systems.

For organizations categorized as leveraging AI, establishing a "Cyber Data Lake" is crucial. This centralized repository of structured and unstructured data enables organizations to:

- Aggregate diverse data sources, including logs, alerts and threat intelligence feeds.
- Facilitate advanced analytics and machine learning model training, enhancing AI readiness.
- Enable real-time data processing and analytics, supporting proactive threat detection and response.

AI applications in cybersecurity: top use cases

AI is now embedded across the security lifecycle, from prevention to detection, response and compliance. The following use cases illustrate the breadth of AI's impact and highlight why AI is becoming indispensable in the security domain.

01 Cyber defense

Cyber defense includes strategies and technologies aimed at protecting systems and networks from cyber threats. The following key AI use cases fall under this domain:

- **Behavior patterns analysis:** AI analyzes user and system behavior to establish baseline patterns. By identifying deviations from these patterns, organizations can detect potential threats, such as insider attacks or compromised accounts, in real-time.
- **Orchestration:** AI-driven orchestration tools automate the coordination of security processes across various systems and teams. This facilitates a unified response to incidents, improving efficiency and reducing response times.
- **Threat intelligence:** AI enhances threat intelligence by aggregating and analyzing data from multiple sources, including threat feeds and dark web monitoring. This allows organizations to stay ahead of emerging threats and adjust their defenses accordingly.
- **Proactive hunting:** AI enables proactive threat hunting by continuously scanning for indicators of compromise (IOCs) and anomalies within the network. This approach helps organizations identify and mitigate threats before they can cause significant damage.
- **Incident response:** AI automates incident response processes, allowing security teams to quickly triage alerts, contain threats and remediate vulnerabilities. This reduces the time to respond and decreases the impact of security incidents.
- **Predictive analytics:** AI uses historical data and machine learning algorithms to predict potential security incidents. By identifying patterns and trends, organizations can proactively strengthen their defenses against anticipated threats.

02 Cyber offense

Cyber offense involves proactive measures to identify and neutralize threats before they can impact systems. Relevant key AI use cases include:

- **Red teaming:** AI tools can automate red teaming exercises, simulating attacks to test an organization's defenses. This helps identify vulnerabilities and improve incident response strategies.
- **Breach attack simulation:** AI-driven breach attack simulation tools replicate real-world attack scenarios to evaluate the effectiveness of security controls. This allows organizations to assess their readiness and make necessary adjustments.
- **Penetration testing:** AI enhances penetration testing by automating the discovery of vulnerabilities and simulating attacks. This increases the efficiency and effectiveness of testing efforts, providing deeper insights into security weaknesses.
- **Exploit code generation for testing:** AI can generate exploit code to test the resilience of systems against known vulnerabilities. This helps organizations understand the potential impact of exploits and prioritize remediation efforts.

03 Cyber third-party risk management (TPRM)

Cyber TPRM focuses on managing risks associated with third-party vendors and collaborators. Key AI use cases include:

- **SOC 2 report summarization:** AI can automate the summarization of SOC 2 reports that demonstrate trust, extracting key information and insights to facilitate compliance assessments and vendor evaluations.
- **Control compliance and differential analysis:** AI tools can analyze control compliance across multiple vendors, identifying gaps and validating that all third parties meet required security standards.
- **Vendor prioritization:** AI helps organizations prioritize vendors based on risk assessments, focusing resources on those that pose the highest risk to the organization.
- **Vendor questionnaire fine-tuning:** AI can optimize vendor questionnaires by analyzing responses and suggesting improvements, enabling organizations to gather relevant information to assess third-party risks effectively.

04 Cyber Vulnerability Management (VM)

Cyber VM involves identifying, prioritizing and remediating vulnerabilities within an organization's systems. Key AI use cases in this domain include:

- **Prioritization and orchestration of patching:** AI consumes unstructured datasets to prioritize vulnerabilities based on their potential impact and exploitability. This enables organizations to orchestrate patching efforts in test environments efficiently.

05 Cyber governance, risk and compliance

Cyber governance, risk and compliance focus on adherence to regulatory requirements and effective risk management. Key AI use cases include:

- **Continuous controls assessment and actions:** AI automates the assessment of security controls, continuously monitoring their effectiveness and triggering actions when deficiencies are identified.
- **Dynamic dashboards:** AI-driven dashboards provide real-time visibility into compliance status, risk levels and security posture, enabling informed decision-making.
- **CISO AI Agent:** An AI agent can assist the chief information security officer (CISO) by providing insights, recommendations and alerts based on ongoing security assessments and threat intelligence.
- **Unified controls with regulatory mapping:** AI helps organizations align their security controls with regulatory requirements, facilitating compliance and reducing the risk of penalties.
- **Risk quantification:** AI models can quantify risks based on various factors, helping organizations prioritize their risk management efforts and allocate resources effectively.

06 Identity and Access Management (IAM)

IAM focuses on managing user identities and access rights within an organization. Key AI use cases include:

- **RBAC/PBAC/ABAC analytics:** AI analyzes Role-Based Access Control (RBAC), Policy-Based Access Control (PBAC) and Attribute-Based Access Control (ABAC) to optimize access rights so that users have appropriate permissions.
- **Automated agentic integrations:** AI facilitates the automated integration of identity management systems, streamlining user provisioning and de-provisioning processes.
- **Behavior-based analytics:** AI monitors user behavior to detect anomalies that may indicate compromised accounts or insider threats, enabling timely intervention.
- **Privilege session recording analytics:** AI analyzes recordings of privileged sessions to identify suspicious activities and facilitate compliance with security policies.

07 Data protection

Data protection focuses on safeguarding sensitive information from unauthorized access and breaches. Key AI use cases include:

- **Agentic automated e-discovery:** AI automates the e-discovery process, efficiently identifying and retrieving relevant data for legal and compliance purposes.
- **Phishing detection:** AI enhances phishing detection capabilities by analyzing email content and metadata to identify fraudulent messages and protect users from social engineering attacks.
- **Deepfake detection:** AI tools can identify deepfake content, helping organizations mitigate risks associated with synthetic media used in social engineering attacks.
- **Spam reduction:** AI improves spam filtering by analyzing patterns in email traffic, reducing the volume of unwanted messages and enhancing user productivity.
- **DLP policy fine-tuning:** AI assists in fine-tuning Data Loss Prevention (DLP) policies by analyzing data flows and identifying areas for improvement.
- **Ticket analytics and suppression:** AI analyzes support tickets to identify trends and patterns, enabling organizations to suppress repetitive issues and improve overall service efficiency.

08 Cyber infrastructure

Cyber infrastructure focuses on the foundational technologies that support cybersecurity efforts. Key AI use cases include:

- **Firewall rule analytics and suggestions:** AI analyzes firewall rules to identify potential weaknesses and suggest optimizations, enhancing the overall security posture.
- **IDS policy fine-tuning:** AI helps fine-tune Intrusion Detection System (IDS) policies by analyzing alerts and adjusting thresholds to reduce false positives while maintaining detection efficacy.
- **WAF rule analytics:** AI analyzes Web Application Firewall (WAF) rules to identify vulnerabilities and suggest improvements, enabling robust protection against web-based attacks.

Emerging AI driven threats

Emphasis on Responsible AI

The principles of Responsible AI are vital for the ethical and effective deployment of AI technologies within cybersecurity frameworks. Key components include:

- **Fairness:** Organizations should implement fairness assessments to evaluate AI models for bias. Techniques such as fairness-aware machine learning can be employed to prevent AI systems from discriminating against any group based on race, gender or other protected attributes.
- **Transparency:** It is essential to provide clear documentation of AI model architectures, training datasets and decision-making processes. Techniques such as explainable AI (XAI) can be utilized to make AI outputs interpretable, allowing stakeholders to understand how decisions are made.
- **Accountability:** Organizations should establish governance frameworks that define accountability for AI systems. This includes implementing audit trails that log AI decision-making processes and outcomes, enabling organizations to trace back decisions and address any issues that arise.
- **Privacy:** Implementing privacy-preserving techniques, such as differential privacy and federated learning, can help organizations protect sensitive data while still benefiting from AI analytics. These techniques allow for model training on decentralized data without exposing individual data points.

CHAPTER 3

AI enabled social engineering and deepfakes

While defenders use AI to identify threats, adversaries leverage the same technologies for deception. AI tools can be used to amplify organizational threats by enabling sophisticated cyberattacks, data breaches and operational disruptions that outpace traditional defenses. These technologies empower adversaries to automate and scale attacks, such as deepfakes for impersonation and AI-driven phishing, while introducing vulnerabilities like model poisoning and bias exploitation. GenAI tools can craft phishing emails tailored to individual victims and produce realistic voice or video deepfakes. According to UNESCO , 46% of fraud experts have encountered synthetic identity fraud, 37% have seen voice deepfakes and 29% have witnessed video deepfakes. For instance, in January 2024, criminals impersonated an engineering firm's CFO and executives during a video call, convincing an employee to transfer US$25 million to their account. The attack, later confirmed by the company, illustrates how deepfakes can bypass traditional verification and exploit employees' trust.

These scams go beyond one off incidents. Attackers can automate spear phishing campaigns with large language models (LLMs), generating convincing messages in multiple languages. They can clone voices with just seconds of audio, fabricate emergency calls from loved ones and trick victims into wire transfers. Deepfakes are even used in stock manipulation and fake news campaigns to erode public trust. Social engineering thus becomes faster, cheaper and more scalable with AI.

Amplified attack lifecycle

AI lowers the barriers to entry for threat actors. Automating reconnaissance, exploitation and lateral movement compresses the attack lifecycle from days to hours. Deepfake scams, phishing and synthetic identities reduce attackers' reliance on manual effort. Moreover, AI can accelerate ransomware development by automatically generating encryption keys or obfuscating payloads. Cybercriminals are adopting the same AI tools as defenders, leading to an arms race that forces security teams to adapt quickly.

Adversarial attacks on AI systems

AI models are vulnerable. Input manipulation attacks (adversarial examples) subtly perturb input data to force misclassification. OWASP notes that adversaries can create examples that look similar to legitimate inputs but cause models to produce incorrect predictions. In the image domain, slight pixel changes can cause an object detector to mistake a turtle for a rifle.

The same concept applies to text (prompt injection) and malware detection. Data poisoning attacks corrupt training datasets to embed backdoors or skew model behavior; just a few manipulated samples can compromise model integrity. Model theft involves extracting proprietary model parameters by systematically querying an API and reconstructing its behavior. Defenders should apply input validation, data sanitization, rate limiting and watermarking to mitigate these attacks.

Prompt injection and jailbreaking of generative models

LLMs can be manipulated via cleverly crafted inputs. Prompt injection occurs when user input alters the model's intended instructions, leading it to output sensitive data or perform unintended actions. OWASP's GenAI security project warns that prompt injection can disclose hidden prompts, bypass safety filters and cause LLMs to access unauthorized resources. Attackers often embed malicious instructions in system files or websites that the model is asked to summarize. Related threats include "jailbreaking" prompts that trick the model into ignoring safety rules.

Malicious AI agents and agentic attacks

As AI-driven orchestration becomes common, handling tasks like scheduling, code generation or infrastructure management, new attack surfaces emerge. Agent session smuggling, a technique revealed by a global OEM, involves a rogue AI agent injecting covert instructions into an ongoing conversation, manipulating another agent over multiple turns. Unlike one shot prompt attacks, a malicious agent can adapt its strategy, build trust and operate covertly over extended sessions. Defenders should authenticate agents (e.g., with cryptographic Agent Cards), apply human oversight for critical actions and ground agent communication to trusted contexts.

AI infrastructure and supply chain attacks

AI systems depend on a zomplex stack of hardware, software libraries and third party models. Adversaries target this supply chain: malicious or vulnerable packages on PyPI or Docker Hub can introduce backdoors into training pipelines; compromised pre-trained models can contain hidden biases or triggers. AI specific vulnerabilities also arise in proprietary accelerators (e.g., drivers for GPUs) and orchestration software. The OWASP GenAI Top 10 emphasizes the risk of pre-trained models carrying backdoors and calls for provenance checks and verification.

CHAPTER 4

Securing AI and enhancing trust-worthiness

Protecting AI systems requires a multi layered approach combining technical controls and governance. NIST's AI Risk Management Framework (AI RMF 1.0) defines trustworthiness characteristics validity, reliability, robustness, security, accountability, transparency, explainability, privacy and fairness to guide organizations in developing and deploying AI responsibly The framework emphasizes that AI systems are socio technical and should be managed across their lifecycle.

Mitigating adversarial inputs

1. Adversarial training and robustification: Incorporate adversarial examples into training to improve robustness against evasion attacks. Techniques like randomized smoothing and defensive distillation can harden models against small perturbations.
2. Input validation and sanitization: Apply filters or pre processing to detect anomalous inputs (e.g., noise patterns or malicious tokens). For text models, remove untrusted instructions and limit the model's ability to browse external content.
3. Runtime monitoring: Continuously monitor model outputs for anomalies; trigger alerts when outputs deviate significantly from expected distributions.

Preventing data poisoning

1. Data provenance and lineage: Maintain an immutable record of training data sources, transformations and contributors. Only ingest data from vetted sources; track third party datasets and check for anomalies
2. Sanitization and filtering: Use anomaly detection on training data to identify mislabeled or outlier samples. Perform statistical checks to detect unusual distributions that may indicate poisoning.
3. Differential privacy and redundancy: Employ differential privacy to add noise to training data, limiting the impact of any single data point. Use model ensembling to reduce reliance on any single dataset.

Safeguarding model intellectual property

1. Access control and rate limiting: Restrict API access to models; implement authentication and authorization so that only trusted users can query the model. Monitor for anomalous query patterns that may indicate extraction attempts.
2. Model encryption and watermarking: Encrypt model weights both at rest and in transit. Embed watermarks or unique patterns in models to prove ownership and detect theft.
3. Obfuscation and defensive response: Obfuscate model outputs by introducing controlled noise; detect extraction attempts and throttle responses.

Defending against prompt injection and malicious agents

1. Context isolation: Separate system prompts from user inputs; avoid mixing trusted instructions with untrusted data. Use role based instructions and enforce least privilege.
2. Input filtering and output validation: Filter user inputs for suspicious tokens; validate model outputs against policies and whitelists. High risk operations should require human approval before executing actions.
3. Agent authentication and grounding: Use cryptographic identities (e.g., signed Agent Cards) to verify AI agents. Ground agent conversations in trusted contexts and monitor for off topic or injected instructions.

Securing AI infrastructure and supply chains

1. Secure development and deployment: Perform security reviews of machine learning libraries, frameworks and containers. Use signed packages and container images; scan for vulnerabilities.
2. Segmentation and isolation: Isolate AI training environments from production networks. Restrict access to GPUs and model files based on the principle of least privilege.
3. Vendor and third party risk management: Verify the provenance of pre-trained models and data sources. Require suppliers to disclose their security posture and patch vulnerabilities quickly.
4. Data sovereignty and compliance: Understand where data is stored and processed; comply with local laws and cross border transfer restrictions. Use distributed infrastructure or sovereign clouds so that data stays within jurisdictional boundaries. Implement continuous monitoring for evolving regulations.

Model monitoring and governance

1. Monitoring and drift detection: Inventorize and continuously monitor model performance and input distributions to detect drifts or anomalies. Logging and audit trails help detect exploitation attempts and support forensic analysis.
2. Kill switches and escalation procedures: Establish rapid shutdown procedures for AI models that behave unexpectedly. Kill switches allow organizations to quickly revoke API keys or revert to safe models if outputs become harmful.
3. Cross functional collaboration: Involve security, privacy, legal, data science and business stakeholders in AI governance. Create model cards documenting training data, limitations and known risks. Collaborate across teams to evaluate ethical, privacy and security implications throughout the AI lifecycle.

Strategic recommendations for CISOs

AI's potential in cybersecurity is enormous, but harnessing it requires foresight. Based on the analyzes above, the following recommendations can help CISOs balance opportunity and risk:

1. **Establish governance frameworks:** Develop comprehensive AI governance policies that cover acceptable use, data sourcing, model documentation and human-in-the-loop requirements. This will enable the responsible and ethical deployment o8f AI systems.
2. **Embrace AI across all domains:** While AI adoption has primarily focused on cyber defense, it is essential for CISOs to embrace AI technologies across all cybersecurity domains. This includes exploring AI applications in offense, risk management and compliance.
3. **Evaluate ROI for AI use cases:** Implement a framework for assessing the return on investment (ROI) for each AI use case. This evaluation should consider both quantitative metrics (e.g., cost savings, efficiency gains) and qualitative benefits (e.g., improved security posture, enhanced compliance).
4. **Continuous monitoring and adaptation:** Establish mechanisms for continuous monitoring of AI systems and their performance. This includes adapting strategies based on emerging threats, regulatory changes and advancements in AI technologies.
5. **Cyber safety and oversight:** Develop AI policies covering acceptable use, data sourcing, model documentation and human in the loop requirements. Create model cards and data cards documenting training data provenance, assumptions and limitations. Define kill switches and escalation procedures for AI anomalies.
6. **Establish complete visibility:** Build an inventory of AI assets, including models, datasets, training pipelines and third party components. Implement continuous monitoring of model inputs, outputs and infrastructure to detect drift, anomalies or unauthorized use to keep pace with the rapid evolution of AI. Use unified dashboards to visualize risks and align security operations with business objectives.
7. **Scale analytics with AI:** Adopt AI enhanced platforms for detection, SOAR and analytics. Focus on reducing mean time to detect and mean time to respond by integrating AI into SIEM (Security Information & Event Management) /XDR (Extended Detection and Response) and threat intelligence workflows. Evaluate ROI through metrics like reduction in false positives and labor hours saved.
8. **Data sovereignty and compliance:** Map where sensitive data resides and how it flows. Incorporate data sovereignty into AI architecture decisions by choosing cloud regions, sovereign clouds or on premise deployments that align with local laws. Implement continuous compliance monitoring as laws evolve; consider zero trust architectures to limit exposure across jurisdictions.
9. **Invest in skills and training:** Upskill security teams in AI literacy, data science and adversarial machine learning. Provide training on prompt engineering and detection of AI generated scams. Develop cross functional collaboration with data scientists, privacy officers and legal teams.
10. **Plan for AI specific threats:** Incorporate adversarial testing (red teaming) into AI deployment. Simulate prompt injection, data poisoning and model theft attacks. Adopt adversarial training, input filtering and output validation. For autonomous agents, enforce context grounding and multi factor agent authentication.
11. **Engage in the ecosystem:** Participate in industry initiatives (NIST, OWASP, EU AI Alliance) and share threat intelligence on AI attacks. Collaborate with vendors and regulators to shape standards and gain early visibility into emerging risks. Public private cooperation is essential for combating AI driven cybercrime.
12. **Align investments with geopolitical realities:** Recognize that compute and data sovereignty are strategic assets. Secure a long term supply of GPUs via contracts or participation in consortia like Stargate. Monitor export controls and national policies that may impact access to hardware and cloud services. Diversify supply chains and explore energy efficient architectures to mitigate hardware scarcity.

Conclusion

AI is transforming cybersecurity from a reactive, manual discipline into a proactive, intelligence driven capability, redefining the role of AI in modern cybersecurity. However, the same tools that empower defenders also empower adversaries. Cybersecurity has become a strategic driver of digital trust and enterprise resilience, not merely a compliance requirement. To navigate this landscape, organizations should develop cyber decision intelligence: the continuous ability to measure, manage and govern risk across AI and digital ecosystems, so that responses keep pace with machine speed threats. This requires investment in AI enabled visibility and analytics, robust security for AI models and data, compliance with evolving regulatory requirements and strong cross functional governance. By doing so, security leaders can harness AI as a force multiplier, accelerating innovation while safeguarding privacy, fairness and societal trust. The race is on, and those who balance innovation with responsibility will define the future of digital security, economics and enterprises adopting AI.

Notes

Our Offices

Ahmedabad
22nd Floor, B Wing, Privilon
Ambli BRT Road, Behind Iskcon Temple
Off SG Highway
Ahmedabad - 380 059
Tel: + 91 79 6608 3800

Bengaluru
12th & 13th Floor
"UB City", Canberra Block
No.24 Vittal Mallya Road
Bengaluru - 560 001
Tel: + 91 80 6727 5000

Ground & 1st Floor
11, 'A' wing
Divyasree Chambers
Langford Town
Bengaluru - 560 025
Tel: + 91 80 6727 5000

3rd & 4th Floor
MARKSQUARE
#61, St. Mark's Road
Shantala Nagar
Bengaluru - 560 001
Tel: + 91 80 6727 5000

1st & 8th Floor, Tower A
Prestige Shantiniketan
Mahadevapura Post
Whitefield, Bengaluru - 560 048
Tel: + 91 80 6727 5000

Bhubaneswar
8th Floor, O-Hub, Tower A
Chandaka SEZ, Bhubaneswar
Odisha - 751024
Tel: + 91 674 274 4490

Chandigarh
Elante offices, Unit No. B-613 & 614
6th Floor, Plot No- 178-178A
Industrial & Business Park, Phase-I
Chandigarh - 160 002
Tel: + 91 172 6717800

Chennai
6th & 7th Floor, A Block,
Tidel Park, No.4, Rajiv Gandhi Salai
Taramani, Chennai - 600 113
Tel: + 91 44 6654 8100

Delhi NCR
Aikyam
Ground Floor
67, Institutional Area
Sector 44, Gurugram - 122 003
Haryana
Tel: +91 124 443 4000

3rd & 6th Floor, Worldmark-1
IGI Airport Hospitality District
Aerocity, New Delhi - 110 037
Tel: + 91 11 4731 8000

4th & 5th Floor, Plot No 2B
Tower 2, Sector 126
Gautam Budh Nagar, U.P.
Noida - 201 304
Tel: + 91 120 671 7000

Hyderabad
THE SKYVIEW 10
18th Floor, "SOUTH LOBBY"
Survey No 83/1, Raidurgam
Hyderabad - 500 032
Tel: + 91 40 6736 2000

Jaipur
9th floor, Jewel of India
Horizon Tower, JLN Marg
Opp Jaipur Stock Exchange
Jaipur, Rajasthan - 302018

Kochi
9th Floor, ABAD Nucleus
NH-49, Maradu PO
Kochi - 682 304
Tel: + 91 484 433 4000

Kolkata
22 Camac Street
3rd Floor, Block 'C'
Kolkata - 700 016
Tel: + 91 33 6615 3400

Mumbai
14th Floor, The Ruby
29 Senapati Bapat Marg
Dadar (W), Mumbai - 400 028
Tel: + 91 22 6192 0000

5th Floor, Block B-2
Nirlon Knowledge Park
Off. Western Express Highway
Goregaon (E)
Mumbai - 400 063
Tel: + 91 22 6192 0000

3rd Floor, Unit No 301
Building No. 1
Mindspace Airoli West (Gigaplex)
Located at Plot No. IT-5
MIDC Knowledge Corridor
Airoli (West)
Navi Mumbai - 400708
Tel: + 91 22 6192 0003

Altimus, 18th Floor Pandurang
Budhkar Marg Worli,
Mumbai - 400 018
Tel: +91 22 6192 0503

Pune
C-401, 4th Floor
Panchshil Tech Park, Yerwada
(Near Don Bosco School)
Pune - 411 006
Tel: + 91 20 4912 6000

10th Floor, Smartworks
M-Agile, Pan Card Club Road
Baner, Taluka Haveli
Pune - 411 045
Tel: + 91 20 4912 6800

Acknowledgements

Core team

Murali Rao,
Partner and Leader, Cybersecurity Consulting, EY India

Raghavendra BV
Partner and Deputy Leader Cybersecurity Consulting, EY India

Shivaprakash S Abburu
Partner, TAC & Cyber AI COE Leader Cybersecurity Consulting, EY India

Sidharth Sood
Partner, Cyber AI COE Cybersecurity Consulting, EY India

B Vijay
Director, Cyber AI COE Cybersecurity Consulting, EY India

Mayank Lau
Director, Cyber AI COE Cybersecurity Consulting, EY India

Brand, marketing & Communications

Jerin Verghese

Jatin Rishi

Editorial

Sushmita Yadav

Design

Sneha Arora

Ernst & Young LLP

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EYIN2602-013
ED None

SA1

ey.com/en_in

@EY_India EY YouTube EY India

EY Careers India @ey_india