MAYER|BROWN

AI SUMMIT 2026

New York

Thought Leadership Materials

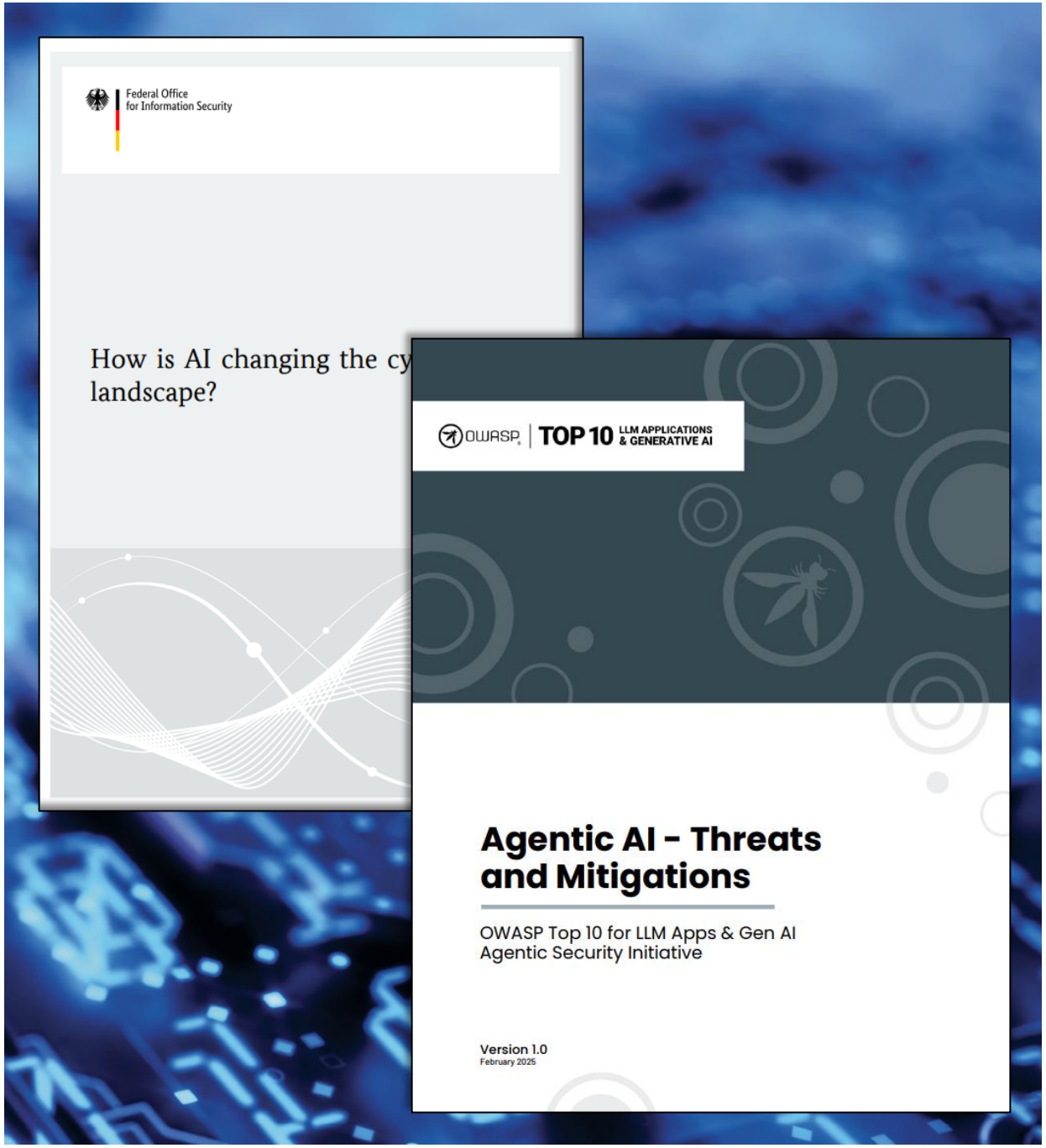
Panel Session 5: Security Challenges:

Managing Risks in the Age of AI

SECURITY CHALLENGES: MANAGING RISKS IN THE AGE OF AI

01

THREATS TO AI



AI SECURITY THREATS

- **Companies face a broad range of attacks on AI systems, including attacks that are common to other software-based systems and attacks that are distinctive to AI systems. Attacks include:**
  - Evasion attacks: malicious input to fool the model or reduce its accuracy, e.g., prompt injection
  - Poisoning attacks, e.g., data poisoning, model poisoning
  - Information extraction attacks, e.g., model stealing, data reconstruction, membership or attribute inference attacks
  - Supply chain attacks, e.g., slopsquatting
  - Abuse of agentic AI
- **Companies also face inadvertent security risks from the use of AI, including from the use of shadow AI or the use of sensitive data in model finetuning or prompts**
- **Companies can turn to an increasing number of resources to understand these threats, such as NIST, OWASP, MITRE Atlas, German BSI.**

02

LEGAL EXPECTATIONS FOR SECURING AI

AI SECURITY BEST PRACTICES WILL INFORM LEGAL EXPECTATIONS FOR COMPANIES

- **Best practices for AI security have been developed along a number of key dimensions of AI security, including:**
  - Data security
  - Application security
  - Model/model weight security
  - Infrastructure security
  - Securing AI output (code development)
- **Companies also face continued—and potentially heightened—expectations to maintain appropriate security for the IT on which AI systems depend.**
- **How exactly these best practices will inform regulatory expectations, litigation claims, and contractual requirements remains to be seen.**

Joint Cybersecurity Information

TLP:CLEAR

National Cyber Security Centre

PART OF THE GCSB

Australian Government

Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE

ACSC

National Cyber Security Centre

a part of GCHQ

AI Data Security

Best Practices for Securing Data Used to Train & Operate AI Systems

Executive summary

This Cybersecurity Information Sheet (CSI) provides essential guidance on securing data used in artificial intelligence (AI) and machine learning (ML) systems. It also highlights the importance of data security in ensuring the accuracy and integrity of AI outcomes and outlines potential risks arising from data integrity issues in various stages of AI development and deployment.

This CSI provides a brief overview of the AI system lifecycle and general best practices to secure data used during the development, testing, and operation of AI-based systems. These best practices include the incorporation of techniques such as data encryption, digital signatures, data provenance tracking, secure storage, and trust infrastructure. This CSI also provides an in-depth examination of three significant areas of data security risks in AI systems: data supply chain, maliciously modified ("poisoned") data, and data drift. Each section provides a detailed description of the risks and the corresponding best practices to mitigate those risks.

This guidance is intended primarily for organizations using AI systems in their operations, with a focus on protecting sensitive, proprietary, or mission critical data. The principles outlined in this information sheet provide a robust foundation for securing AI data and ensuring the reliability and accuracy of AI-driven outcomes.

This document was authored by the National Security Agency's Artificial Intelligence Security Center (AISC), the Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC), the New Zealand's Government Communications

This information is marked TLP:CLEAR. TLP:CLEAR information may be distributed without restriction. For more information on the Traffic Light Protocol, see cisa.gov/tlp/.

U/OO/157249-25 | PP-25-2301 | May 2025 Ver. 1.0

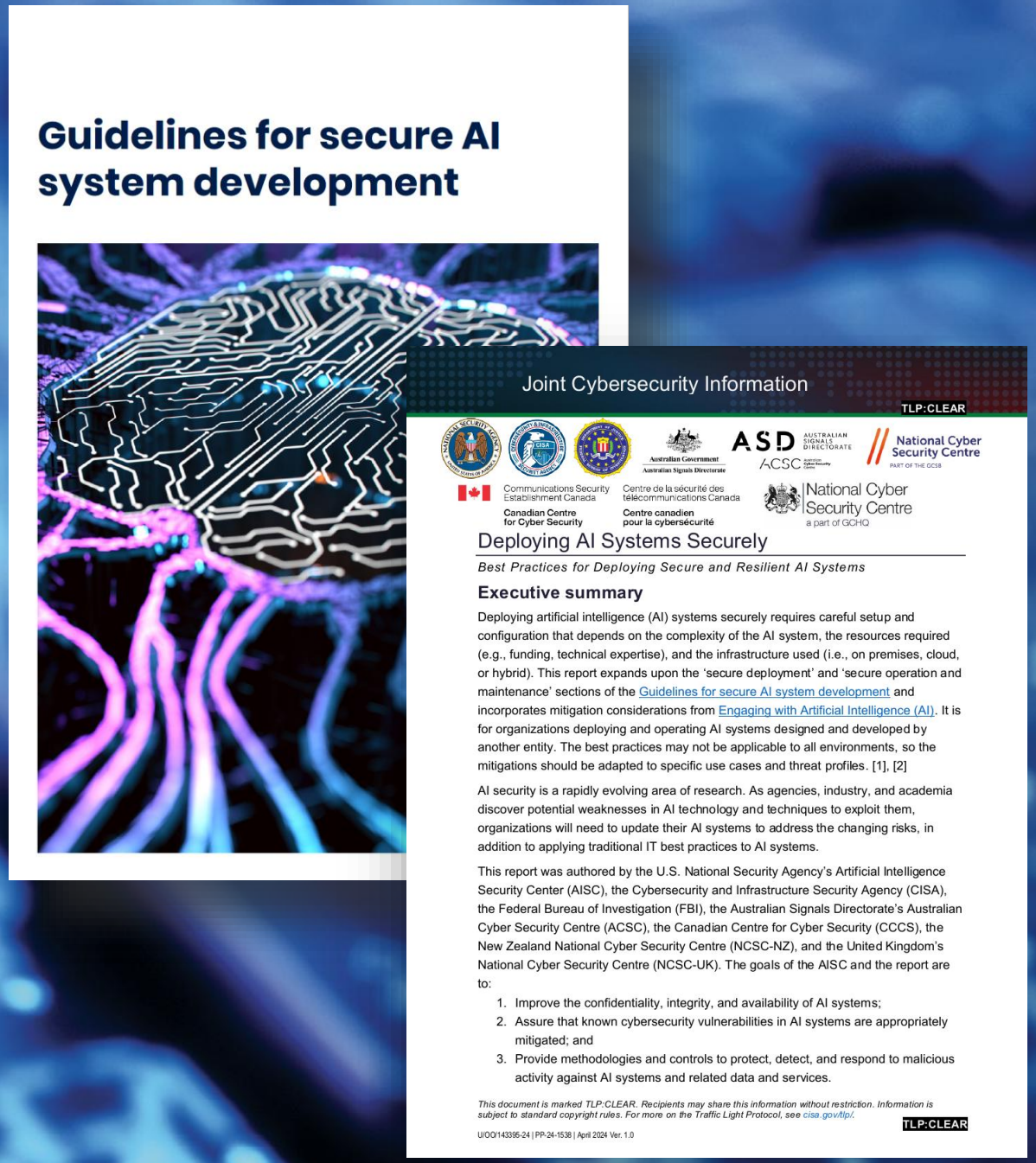
TLP:CLEAR

LEGAL RISKS ARE SIGNIFICANT DESPITE LIMITED SPECIFIC LEGAL REQUIREMENTS

- **The EU AI Act provides limited guidance on security expectations:**
  - Security at system level, but taking into the account other dimensions
  - Guiding principles:
    - Compliance at a system level
    - Security risk assessments needed
    - Integrated and continuous approach
    - Limits in the state of the art for securing AI models
- **However, there are numerous legal frameworks, including many that are sector specific, that inform legal expectations for AI security.**

03

IMPLEMENTING A RISK-BASED AI SECURITY PROGRAM



IMPLEMENTING A RISK-BASED AI SECURITY PROGRAM WILL HELP A COMPANY CAPTURE THE BENEFITS OF AI ADOPTION

- **General cyber risk measures**
  - Threat modeling, risk assessment, and vulnerability testing
  - Strong access controls, identity management, and permission management (e.g. principle of least privilege)
  - Supply chain security and component provenance
  - Logging, monitoring, and incident response planning
- **AI-specific measures**
  - Data provenance, integrity, and bias assessment for training data
  - Adversarial testing, red teaming, and guardrails for prompt injection
  - Monitoring for model drift, data poisoning, and misuse
  - Documentation of model limitations, intended use, and failure modes
- **Key areas to consider when implementing an AI security program include:**
  - Governance
  - Procurement
  - Policies and controls
  - Security testing

EFFECTIVE GOVERNANCE CAN REDUCE RISKS ASSOCIATED WITH AI SECURITY

- Poor calibration of AI security can have significant consequences for a company, whether because it prevents the company from innovating at the necessary pace or because it exposes the company to excessive risk that undermines the benefits of that innovation
- The security team will be an important voice in determining how to manage AI security risk, but this issue will also implicate the expertise and interest of relevant business units, legal, and other stakeholders.
- As in other AI contexts, an effective governance mechanism will help the company appropriately manage AI security risks. This governance will be most effective if it:
  - Includes appropriate stakeholders;
  - Is informed by appropriate risk assessments;
  - Has full visibility into AI deployments across the company;
  - Has authority to impose necessary security measures and processes;
  - Can guide investment decisions into AI-specific security tools;
  - Is implemented through appropriate policies, controls, and procedures;
  - Allows effective executive oversight and decision-making of AI security.

KEY QUESTIONS

- Are necessary stakeholders engaged in managing AI security?
- Does AI security governance fit with other governance mechanisms (e.g., AI, security more broadly)?
- Does AI security governance reach from technical controls to executive decision-making?

RED FLAGS

- Security team is not included in AI governance mechanism
- Development team can disregard security considerations.

FOCUSING ON SECURITY IN AI PROCUREMENT CAN SUBSTANTIALLY REDUCE RISK

- **The procurement process can highlight the potential tension between innovation through rapid adoption of AI tools and ensuring appropriate security that allows the company to fully benefit from that innovation.**
- **Focus on third-party risk**
  - Heightened emphasis on third party risk in recent cyber regulations
  - Particularly relevant in AI context: many layers in supply chain
- **Considerations for procurement teams and their counsel**
  - Take time to understand product and security risk
  - Include AI-specific questions on vendor questionnaires
  - Assess need for security-specific terms to address AI security in contracts with AI vendors
  - Consider impact on other terms, like breach notification, liability

KEY QUESTIONS

- What level of risk does the service provided by the vendor present to the organization?
- Does the vendor meet prevailing security best practices relevant to the service it provides?
- Will the vendor agree to security provisions appropriate to the risk presented by its service ?

RED FLAGS

- Vendor lacks appropriate security maturity.
- Scope of service is unclear or could expand over time.

APPROPRIATE POLICIES AND CONTROLS CAN REDUCE AI SECURITY RISK

- Security policies and controls are likely to vary based on the nature of the company's business and the AI use case, including the sensitivity of the data it will access and the scope of actions it can trigger/take.
- As a baseline, the security policies and controls that apply to other software-based systems presumptively should apply to AI systems to the extent feasible.
- Key issues for attention include: (1) AI tool permissions, for data access and permitted actions; (2) user access rights; (3) system logging and monitoring; (4) data loss prevention; and (5) integration of security into AI development activities.
- With AI-specific security solutions proliferating in the market, security controls should be increasingly automated – and security practices should avoid over-reliance on guidelines for user behavior.
- Companies may wish to update their security policies to address the use of AI or to create specific AI security policies or processes.

KEY QUESTIONS

- Are security policies and controls based on an appropriate assessment of relevant risks?
- Are AI systems built on a weak foundation in that relevant infrastructure lacks appropriate controls?
- Do security controls leverage available technological solutions in an effective way?

RED FLAGS

- AI is implemented with a deploy-first, secure-later mindset
- Paper security policies do not match implemented controls

TESTING THE SECURITY OF AI SYSTEMS WILL HELP ONGOING RISK MITIGATION ACTIVITIES

- **In addition to more traditional security testing, AI red-teaming has important distinctive elements:**
  - Involves adversarial testing methods, e.g., attempts to elicit unwanted behaviors, subvert the model's built-in defenses or guardrails
  - Context-dependent: Red-teaming practices and objectives vary by stakeholder (e.g., commercial developers vs. national security organizations) and by model type (general-purpose vs. specialized models)
- **Challenges:**
  - Measurement: what does it mean to "break" a model, and what constitutes a model failure or vulnerability?
  - Testing across multiple models and tracking results over time
  - Building consensus around testing practices and maintaining transparency
- **Particular questions for frontier models**

KEY QUESTIONS

- Does AI red-teaming account for the distinctive risks associated with AI systems?
- Should the testing be performed at the direction of counsel and the reports subject to legal privilege?
- Are test results incorporated into relevant risk assessments so that they can prioritized along with other key risks?

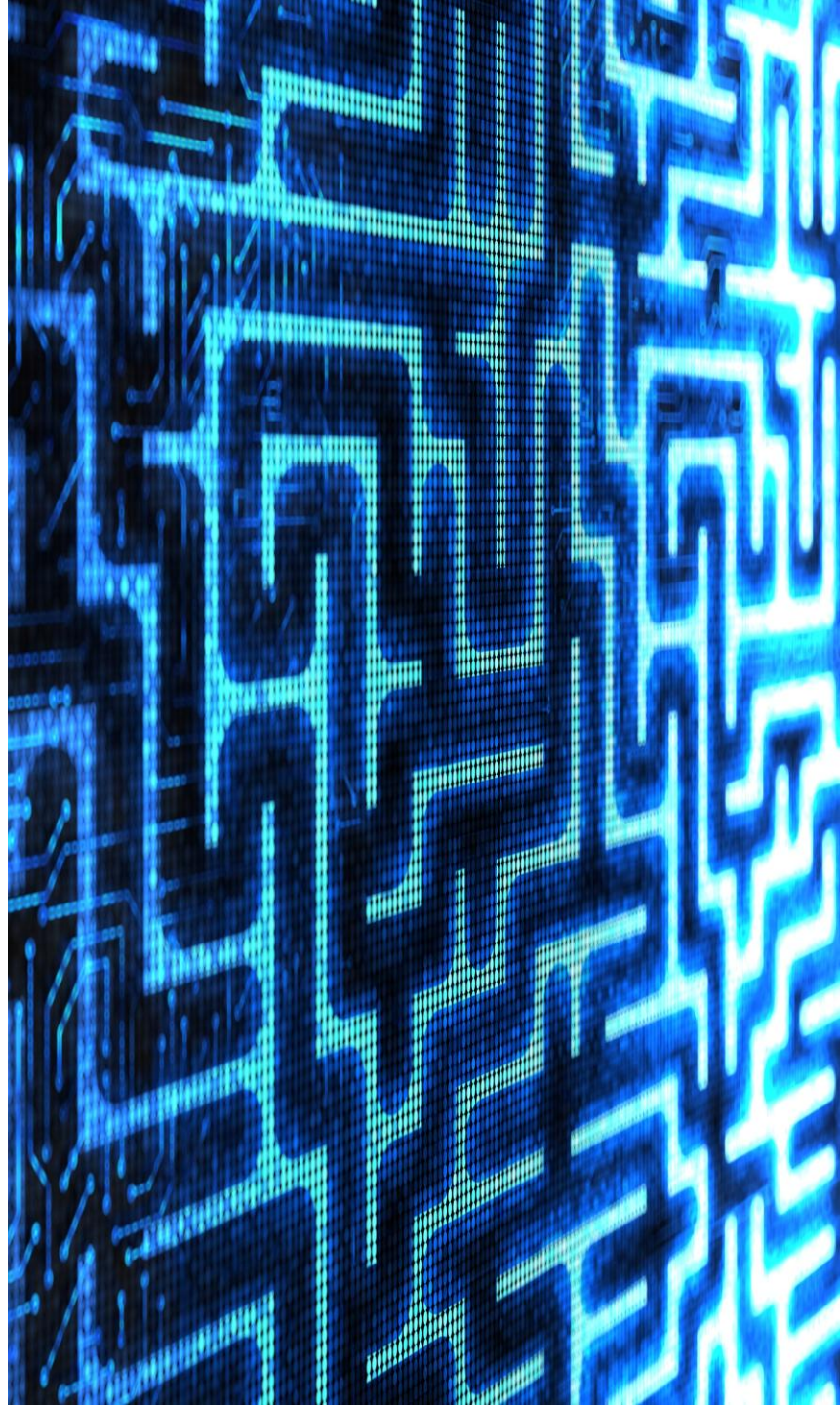
RED FLAGS

- Red-teaming is not tailored to the particular circumstances
- Red-teaming does not inform decision making in a practical way

MAYER | BROWN

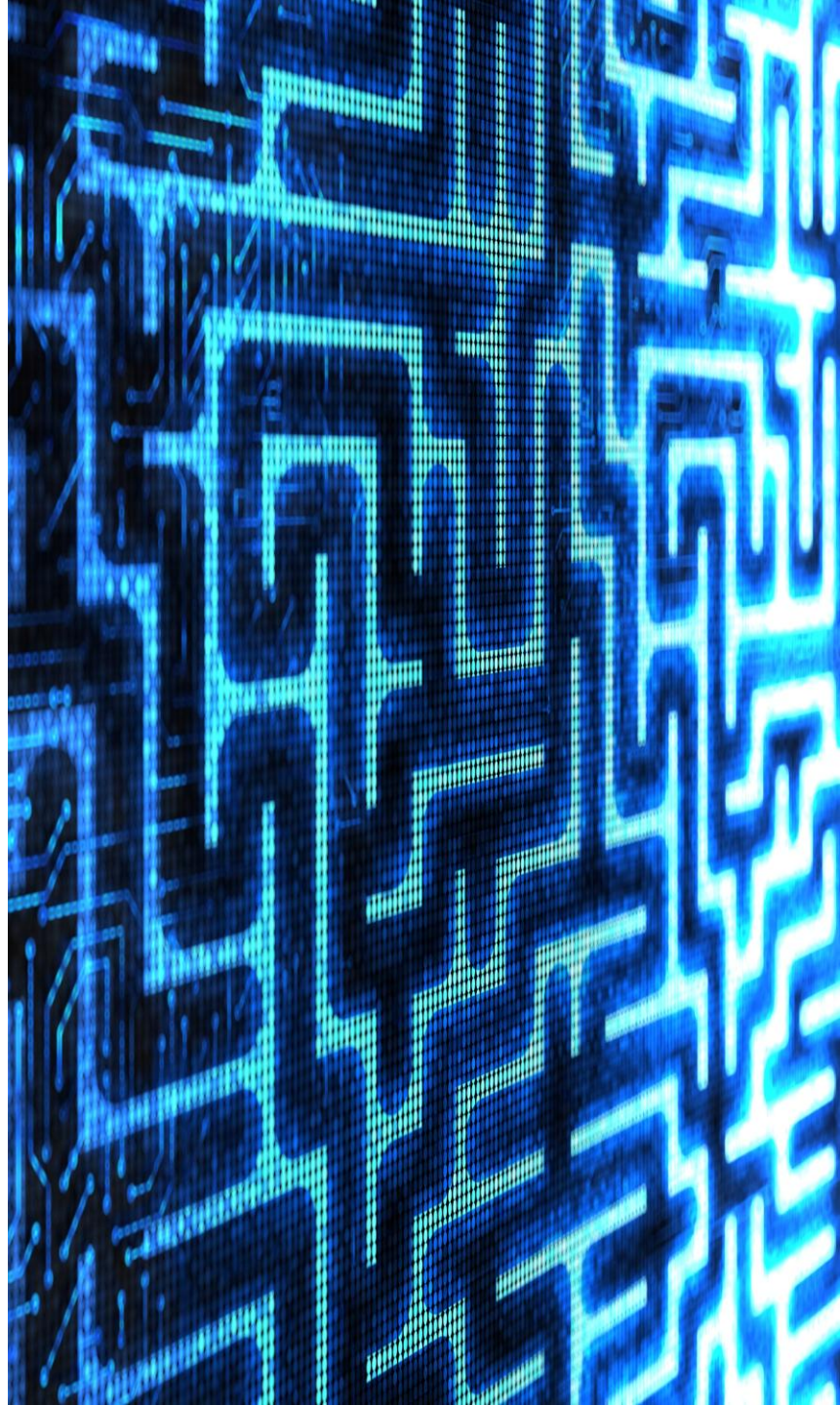
SCENARIO 1: SHADOW AI USE

- Alpha Corp. reviewed a tool provided by an AI vendor called Bad Things Happen AI (“BTH AI”). Alpha Corp. did not approve the tool for three primary reasons: (1) lack of appropriate security controls; (2) inadequate logging of user activity; and (3) insistence by BTH AI that it be allowed to train future models on any data uploaded to the tool.
- Two months later, a security analyst identifies evidence that a member of the human resources team has been using the BTH AI tool.
- When interviewed, the sales team member acknowledges using the tool to analyze certain team members’ employment histories. They can’t remember if they uploaded personal information into the tool. They say that they can no longer access their account after someone changed their password without their permission.



SCENARIO 2: HIGH-VELOCITY AI-POWERED ATTACK

- Beta Corp.'s security team has identified malicious activity in the network.
- Numerous alerts within the security system have flagged multiple accounts escalating privileges, accessing and staging sensitive data, and connecting to suspicious external systems.
- The CISO believes that they have a 5-10 minute window to shut down the relevant systems before the attack is completed.
- The CISO believes that containing the attack would require them to shut off a range of critical systems with unknown commercial and legal consequences.
- The CISO has a 70-80% confidence level that these steps are necessary but they are unsure if they have documented authority to take this step.



SCENARIO 3: AI AGENT EXCEEDS INTENDED AUTHORITY

- Gamma Inc. has recently begun deploying AI agents within its logistics systems. The various agents are subject to a range of controls that are intended to confine their activities to defined areas of authority.
- One of the AI agents has the authority to make changes to route plans among facilities and to enable surge capacity based on upcoming orders, historical trends, and other data sources.
- Rollout of the AI agents has been a significant commercial and public-relations victory, with a reported 3-4% gain in efficiency across key metrics.
- After recent shortfalls at multiple plants, however, the security team investigates and finds that the agent had abused write access to an ordering system to reduce or delete certain orders.