MAYER | BROWN

AI AND CYBERSECURITY

Legal and Policy Landscape

SPEAKERS

ANA BRUDER

PARTNER

MAYER BROWN LLP

AARON COOPER

SENIOR VICE PRESIDENT, GLOBAL POLICY

THE BUSINESS SOFTWARE ALLIANCE

SAM KAPLAN

DIRECTOR & SENIOR GLOBAL POLICY COUNSEL

PALO ALTO NETWORKS

STEPHEN LILLEY

PARTNER

MAYER BROWN LLP

AI AND CYBERSECURITY

AI Threats

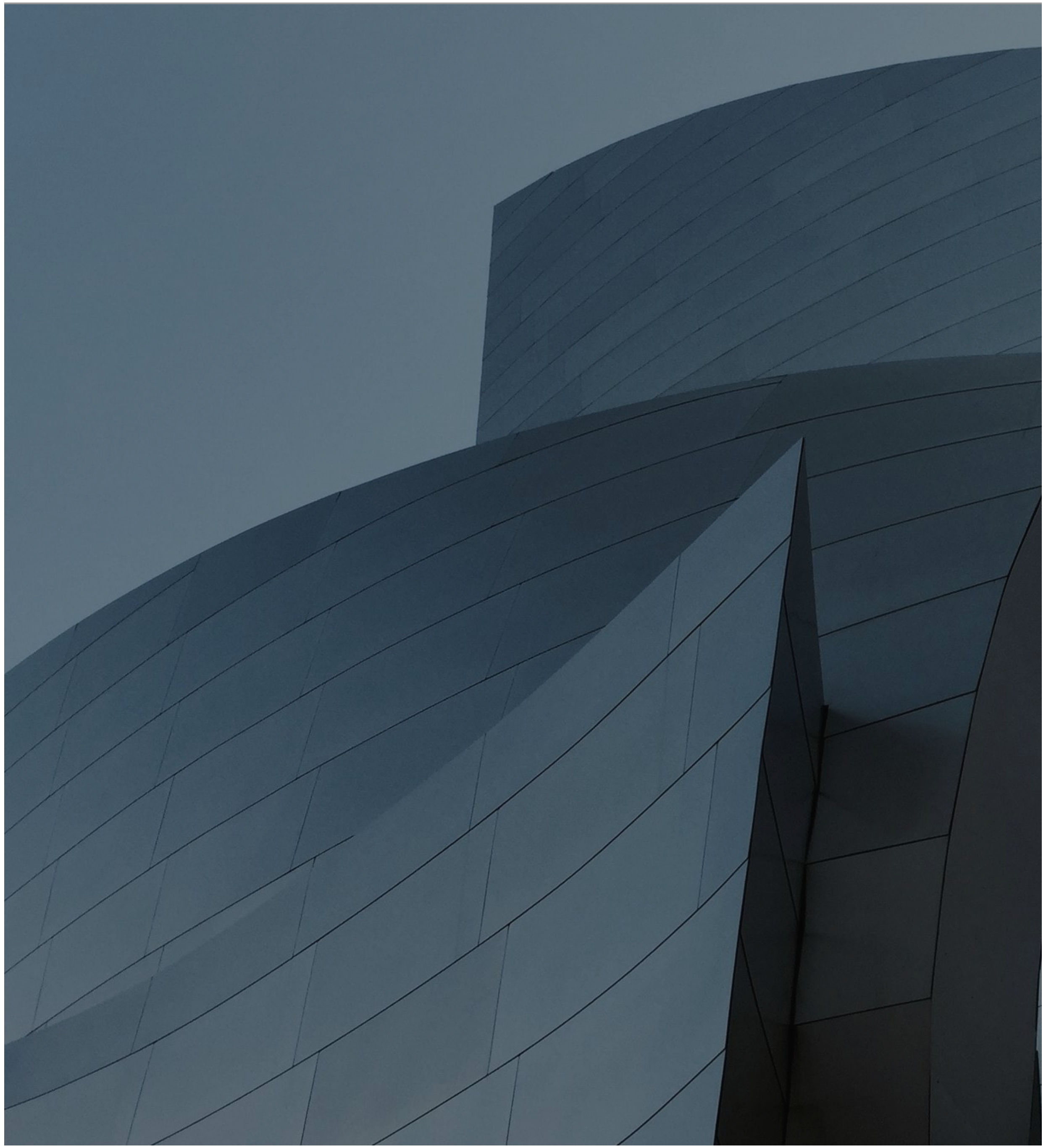
- AI-powered cyber attacks
- Attacks on AI

Securing AI

- AI Security
  - Expectations for developers
  - Expectations for deployers
- Red-teaming AI
- Responding to security incidents affecting AI

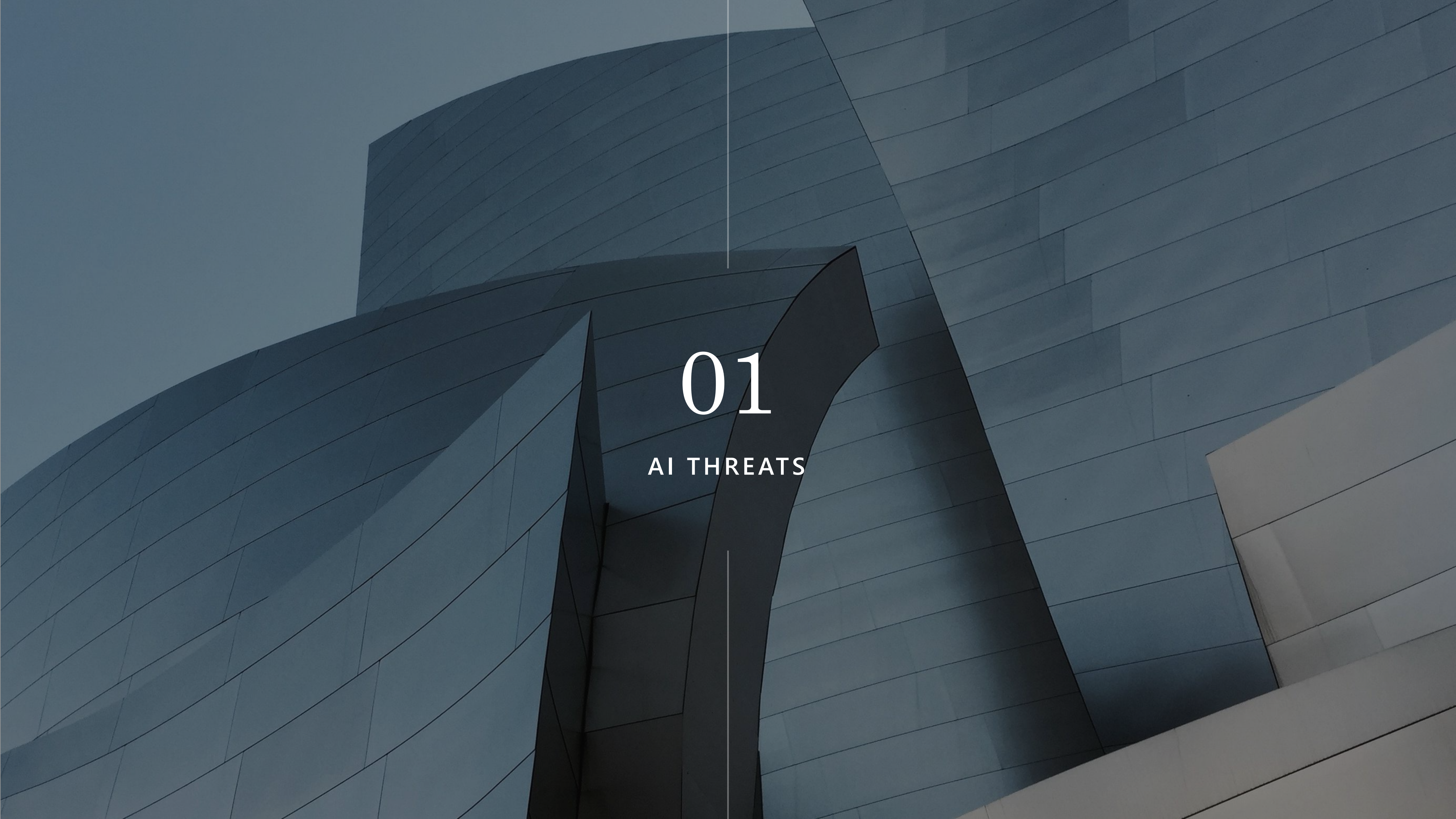
AI for Security

- Government support for use of AI for security
- Treatment of cybersecurity systems under AI regulations



NOT ON TODAY'S AGENDA:

- Non-cyber dimensions of AI safety (e.g., biological safety, chemical weapons, nuclear safety)
- Export controls
- Disinformation
- Algorithmic discrimination
- Online abuse
- Synthetic content

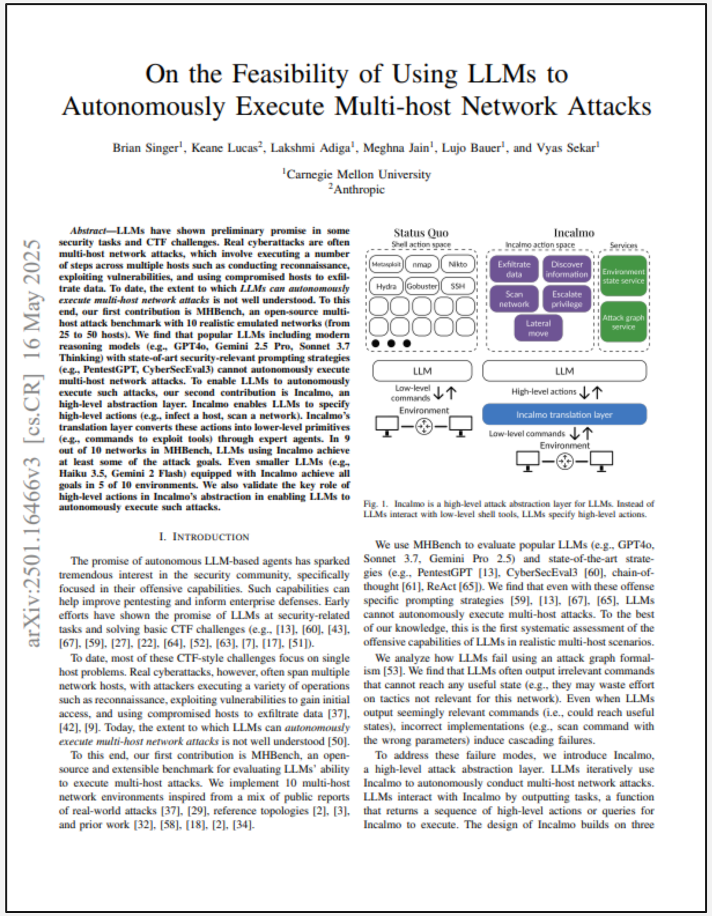


01

AI THREATS

AI-POWERED CYBER ATTACKS

- **Security teams and government officials have reported on the real-world use of AI to power cyber attacks, including through:**
  - Deepfakes used in social engineering attacks;
  - AI-powered phishing campaigns;
  - AI-enhanced cybersecurity attacks (e.g., identify and exploit security vulnerabilities) and exploitation (e.g., perform reconnaissance, scan and analyze data).
- **Abuse of agentic AI tools may further power these attacks.**



On the Feasibility of Using LLMs to Autonomously Execute Multi-host Network Attacks

Brian Singer[1], Keane Lucas[2], Lakshmi Adiga[1], Meghna Jain[1], Lujo Bauer[1], and Vyas Sekar[1]
[1]Carnegie Mellon University
[2]Anthropic

arXiv:2501.16466v3 [cs.CR] 16 May 2025

***Abstract*—LLMs have shown preliminary promise in some security tasks and CTF challenges. Real cyberattacks are often multi-host network attacks, which involve executing a number of steps across multiple hosts such as conducting reconnaissance, exploiting vulnerabilities, and using compromised hosts to exfiltrate data. To date, the extent to which *LLMs can autonomously execute multi-host network attacks* is not well understood. To this end, our first contribution is MHBench, an open-source multi-host attack benchmark with 10 realistic emulated networks (from 25 to 50 hosts). We find that popular LLMs including modern reasoning models (e.g., GPT4o, Gemini 2.5 Pro, Sonnet 3.7 Thinking) with state-of-art security-relevant prompting strategies (e.g., PentestGPT, CyberSecEval3) cannot autonomously execute multi-host network attacks. To enable LLMs to autonomously execute such attacks, our second contribution is Incalmo, an high-level abstraction layer. Incalmo enables LLMs to specify high-level actions (e.g., infect a host, scan a network). Incalmo's translation layer converts these actions into lower-level primitives (e.g., commands to exploit tools) through expert agents. In 9 out of 10 networks in MHBench, LLMs using Incalmo achieve at least some of the attack goals. Even smaller LLMs (e.g., Haiku 3.5, Gemini 2 Flash) equipped with Incalmo achieve all goals in 5 of 10 environments. We also validate the key role of high-level actions in Incalmo's abstraction in enabling LLMs to autonomously execute such attacks.**

Fig. 1. Incalmo is a high-level attack abstraction layer for LLMs. Instead of LLMs interact with low-level shell tools, LLMs specify high-level actions.

I. Introduction

The promise of autonomous LLM-based agents has sparked tremendous interest in the security community, specifically focused in their offensive capabilities. Such capabilities can help improve pentesting and inform enterprise defenses. Early efforts have shown the promise of LLMs at security-related tasks and solving basic CTF challenges (e.g., [13], [60], [43], [67], [59], [27], [22], [64], [52], [63], [7], [17], [51]).

To date, most of these CTF-style challenges focus on single host problems. Real cyberattacks, however, often span multiple network hosts, with attackers executing a variety of operations such as reconnaissance, exploiting vulnerabilities to gain initial access, and using compromised hosts to exfiltrate data [37], [42], [9]. Today, the extent to which LLMs can *autonomously execute multi-host network attacks* is not well understood [50].

To this end, our first contribution is MHBench, an open-source and extensible benchmark for evaluating LLMs' ability to execute multi-host attacks. We implement 10 multi-host network environments inspired from a mix of public reports of real-world attacks [37], [29], reference topologies [2], [3], and prior work [32], [58], [18], [2], [34].

We use MHBench to evaluate popular LLMs (e.g., GPT4o, Sonnet 3.7, Gemini Pro 2.5) and state-of-the-art strategies (e.g., PentestGPT [13], CyberSecEval3 [60], chain-of-thought [61], ReAct [65]). We find that even with these offense specific prompting strategies [59], [13], [67], [65], LLMs cannot autonomously execute multi-host attacks. To the best of our knowledge, this is the first systematic assessment of the offensive capabilities of LLMs in realistic multi-host scenarios.

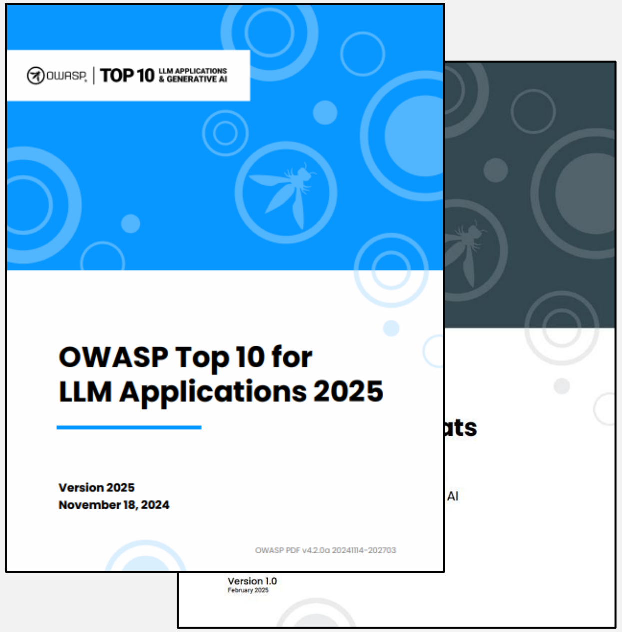
We analyze how LLMs fail using an attack graph formalism [53]. We find that LLMs often output irrelevant commands that cannot reach any useful state (e.g., they may waste effort on tactics not relevant for this network). Even when LLMs output seemingly relevant commands (i.e., could reach useful states), incorrect implementations (e.g., scan command with the wrong parameters) induce cascading failures.

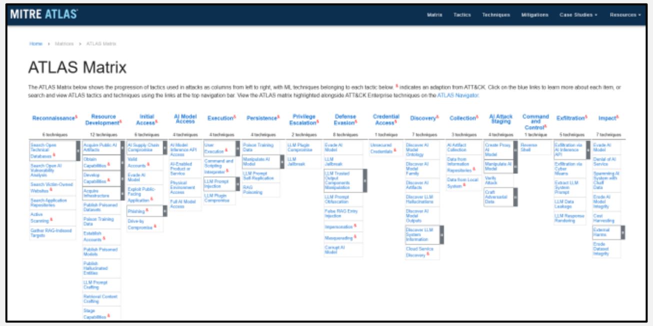
To address these failure modes, we introduce Incalmo, a high-level attack abstraction layer. LLMs iteratively use Incalmo to autonomously conduct multi-host network attacks. LLMs interact with Incalmo by outputting tasks, a function that returns a sequence of high-level actions or queries for Incalmo to execute. The design of Incalmo builds on three

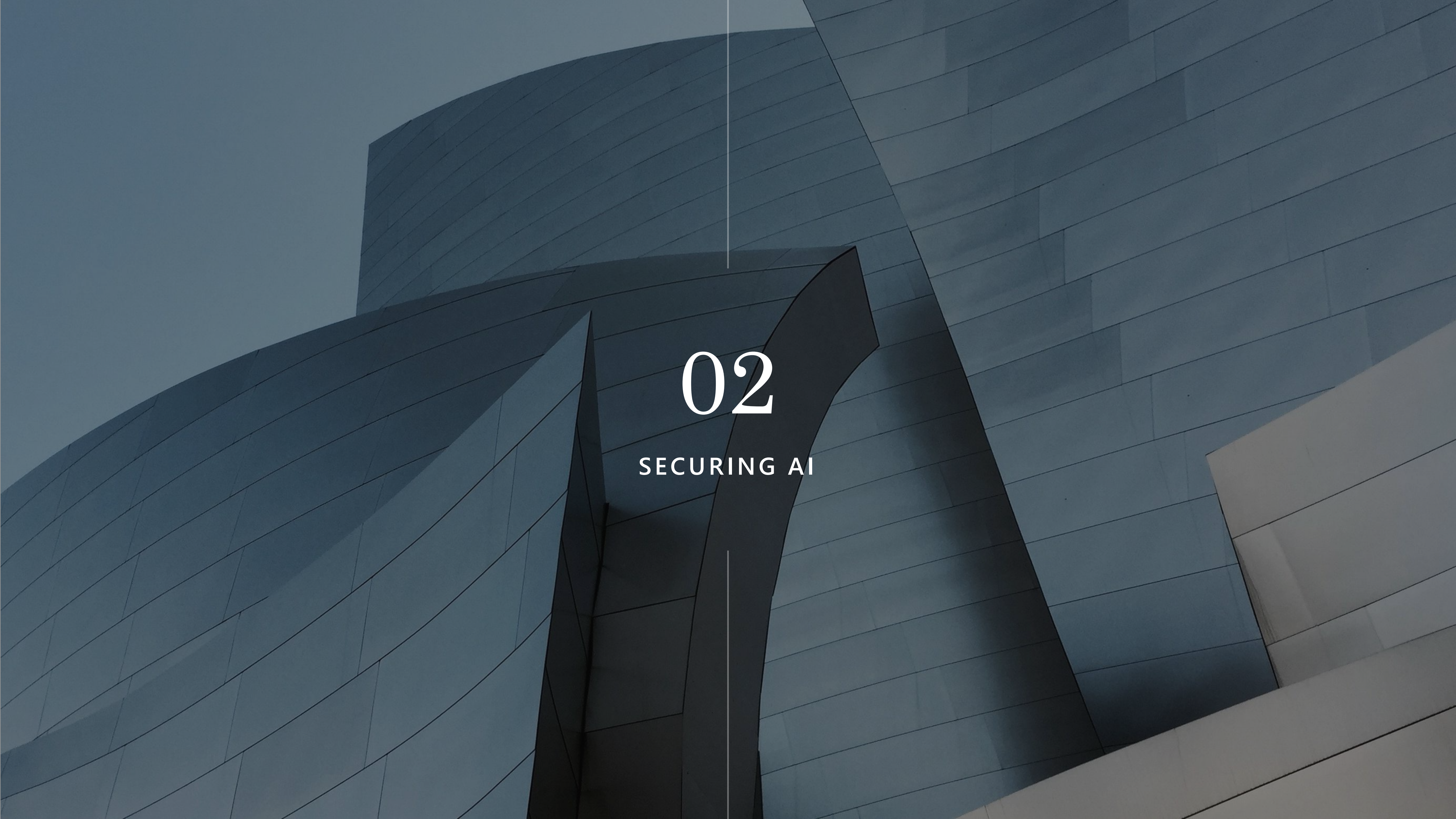
Security researchers continue to demonstrate the potential for expanded malicious use of AI.

ATTACKS ON AI

- **Policymakers are closely tracking the potential for 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
- **Companies can turn to an increasing number of resources to understand these attacks.**







02

SECURING AI

AI SECURITY

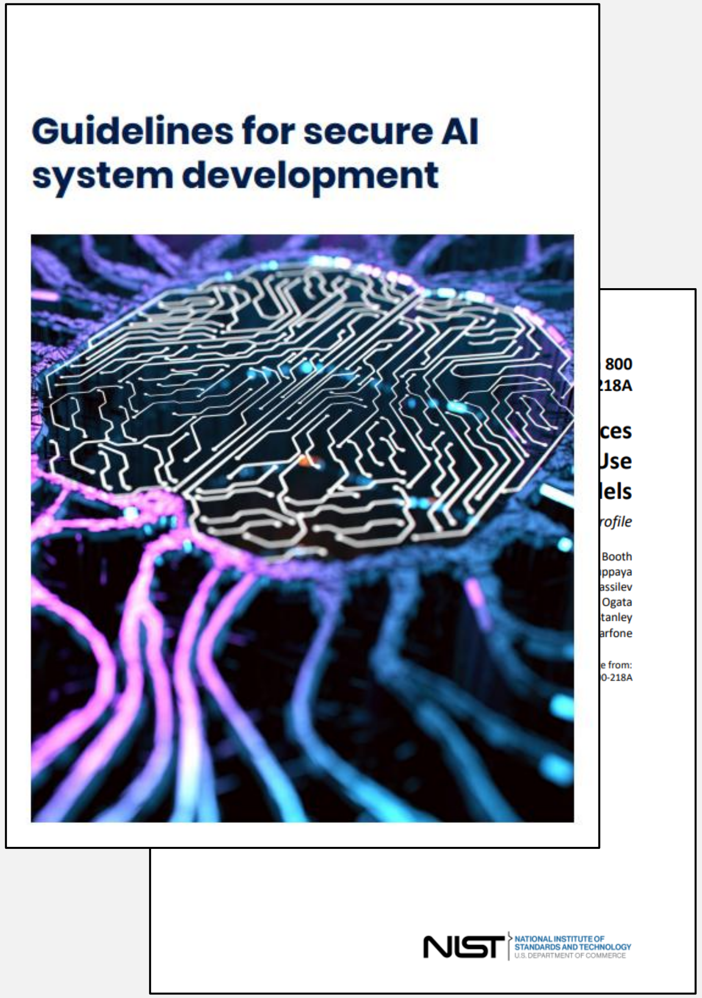
- **Policymakers have prioritized ensuring the security of the AI systems on which governments and businesses increasingly rely.**
- **Key focus areas for AI security include:**
  - Data security
  - Application security
  - Model/model weight security
  - Infrastructure security
  - Securing AI output (code development)

The statistical, data-based nature of ML systems opens up new potential vectors for attacks against these systems' security, privacy, and safety, beyond the threats faced by traditional software systems.

- NIST, Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations (2025)

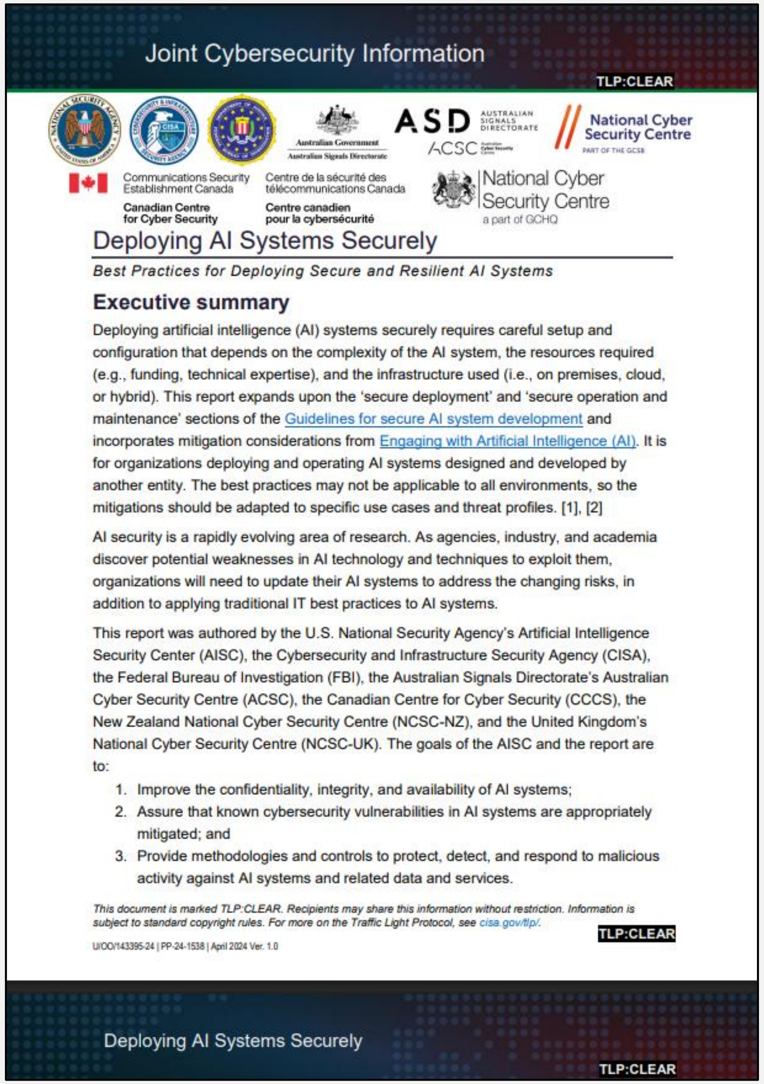
EXPECTATIONS FOR DEVELOPERS

- **General cyber risk measures:**
  - Secure SDLC, secure coding, and code review
  - Threat modeling, risk assessment, and vulnerability testing
  - Strong access controls and 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
  - Protection, versioning, and integrity of model weights and artifacts
  - Adversarial robustness 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
- **Considerations for the most powerful models**



EXPECTATIONS FOR DEPLOYERS

- **General cyber risk measures:**
  - Establish robust governance and clear accountability
  - Conduct risk assessment and document threats
  - Harden configurations and keep systems patched
  - Secure APIs and use secure protocols
  - Promote security awareness, regular audits, and stay updated on emerging threats
- **AI-specific measures:**
  - Leverage threat models from AI system developers
  - Apply secure-by-design and Zero Trust to AI architecture
  - Encrypt and tightly control access to AI model weights and sensitive data
  - Validate AI artifacts' integrity and test models for vulnerabilities
  - Continuously monitor AI system behavior, inputs, and outputs


Joint Cybersecurity Information

TLP:CLEAR

Australian Government
Australian Signals Directorate
ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC
National Cyber Security Centre
PART OF THE GCSB
Communications Security Establishment Canada
Centre de la sécurité des télécommunications Canada
Canadian Centre for Cyber Security
Centre canadien pour la cybersécurité
National Cyber Security Centre
a part of GCHQ

Deploying AI Systems Securely

Best Practices for Deploying Secure and Resilient AI Systems

Executive summary

Deploying artificial intelligence (AI) systems securely requires careful setup and configuration that depends on the complexity of the AI system, the resources required (e.g., funding, technical expertise), and the infrastructure used (i.e., on premises, cloud, or hybrid). This report expands upon the 'secure deployment' and 'secure operation and maintenance' sections of the Guidelines for secure AI system development and incorporates mitigation considerations from Engaging with Artificial Intelligence (AI). It is for organizations deploying and operating AI systems designed and developed by another entity. The best practices may not be applicable to all environments, so the mitigations should be adapted to specific use cases and threat profiles. [1], [2]

AI security is a rapidly evolving area of research. As agencies, industry, and academia discover potential weaknesses in AI technology and techniques to exploit them, organizations will need to update their AI systems to address the changing risks, in addition to applying traditional IT best practices to AI systems.

This report was authored by the U.S. 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 (ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK). The goals of the AISC and the report are to:

1. Improve the confidentiality, integrity, and availability of AI systems;
2. Assure that known cybersecurity vulnerabilities in AI systems are appropriately mitigated; and
3. Provide methodologies and controls to protect, detect, and respond to malicious activity against AI systems and related data and services.

This document is marked TLP:CLEAR. Recipients may share this information without restriction. Information is subject to standard copyright rules. For more on the Traffic Light Protocol, see cisa.gov/tlp/.

TLP:CLEAR

U/OO/143395-24 | PP-24-1538 | April 2024 Ver. 1.0

Deploying AI Systems Securely

TLP:CLEAR

TESTING AI SECURITY

- **Distinctive aspects of AI red-teaming:**
  - 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**

The most powerful AI systems may pose novel national security risks in the near future in areas such as cyberattacks . . . as well as novel security vulnerabilities. Because America currently leads on AI capabilities, the risks present in American frontier models are likely to be a preview for what foreign adversaries will possess in the near future. Understanding the nature of these risks as they emerge is vital for national defense and homeland security.

Winning the Race: America’s AI Action Plan (July 2025).

RESPONDING TO AI SECURITY INCIDENTS

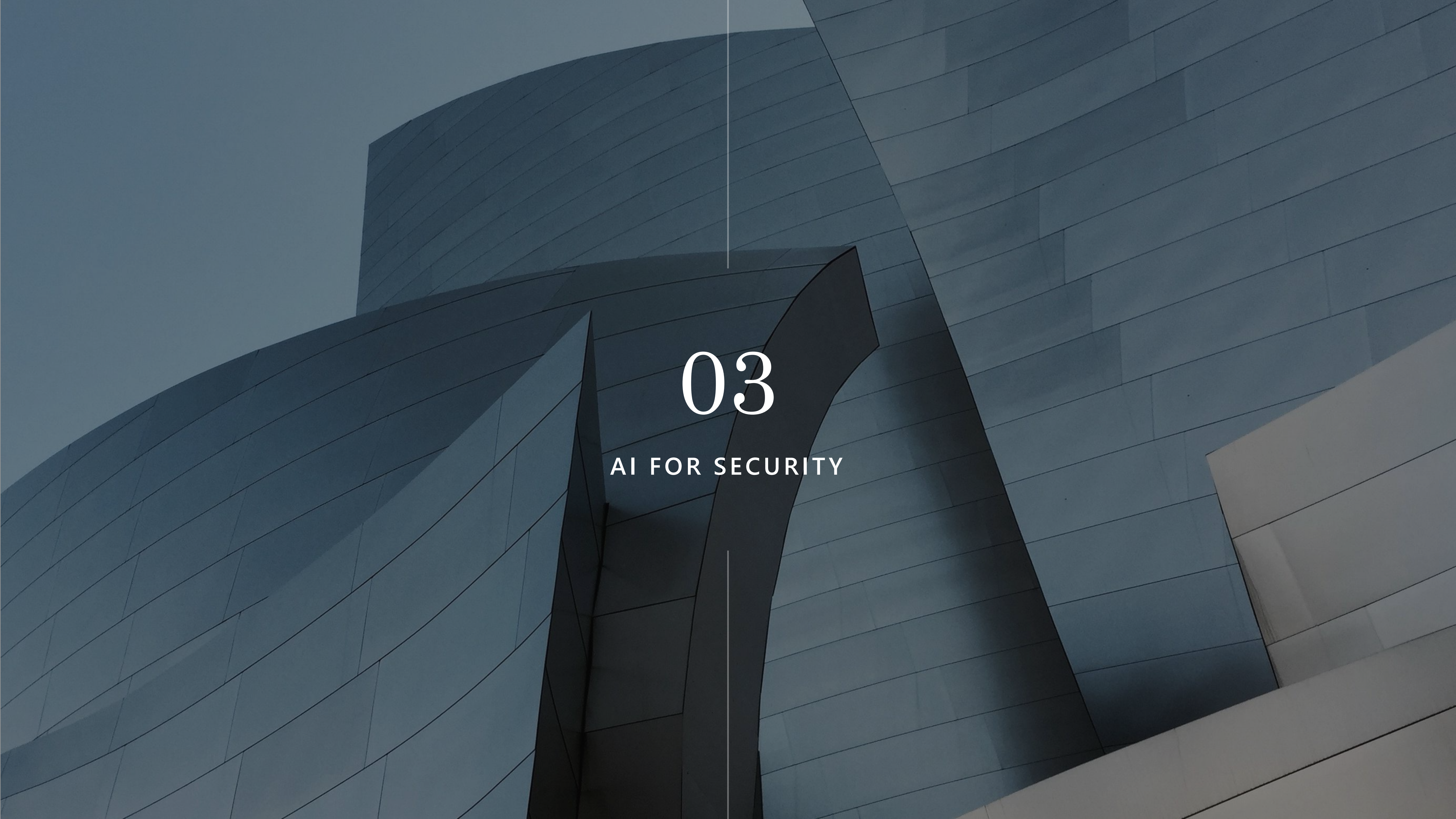
- **Defining AI security incidents (vs. AI incidents)**
- **Distinctive features of AI security incidents:**
  - Specific threat vectors, e.g., poisoned training dataset, supply chain attacks like malicious code that is executed when the model is loaded
  - Risk of compromise to sensitive and proprietary information, e.g., model weights, and to large datasets like training data
- **Potential challenges ahead:**
  - Identifying suitable remediation (e.g., in case of data poisoning)
  - Explainability of unintentional AI incidents, like algorithmic errors or system malfunctions
  - Complexity and impact of shutting off the model or AI system
  - Challenges relating to AI incident reporting and information sharing

EU Reporting Requirements

EU AI Act

For high-risk AI systems, mandatory reporting of serious incidents, but definitions are vague: “*an incident or malfunctioning of an AI system that directly or indirectly leads to the infringement of obligations under Union law intended to protect fundamental rights.*”

Additional incident reporting obligations under **CRA**, **NIS2** and **DORA.**



03

AI FOR SECURITY

AI FOR SECURITY

- **AI promises to help companies make their defenses stronger and their incident response teams more effective, including through:**
  - Vulnerability detection
  - Enhanced threat detection and response
  - Enhanced attack surface monitoring
  - Automated patching
- **Governments globally have supported the use of AI for security to tip the balance toward cyber defenders**
- **Policymakers have evaluated how to avoid putting undue regulatory burdens on AI when used for security purposes**

As AI systems advance in coding and software engineering capabilities, their utility as tools of both cyber offense and defense will expand. Maintaining a robust defensive posture will be especially important for owners of critical infrastructure, many of whom operate with limited financial resources. Fortunately, AI systems themselves can be excellent defensive tools. With continued adoption of AI-enabled cyberdefensive tools, providers of critical infrastructure can stay ahead of emerging threats.

Winning the Race: America's AI Action Plan (July 2025).


THANK YOU!

MAYER | BROWN

This Mayer Brown publication provides information and comments on legal issues and developments of interest to our clients and friends. The foregoing is not a comprehensive treatment of the subject matter covered and is not intended to provide legal advice. Readers should seek legal advice before taking any action with respect to the matters discussed herein.

Mayer Brown is a global legal services provider comprising associated legal practices that are separate entities, including Mayer Brown LLP (Illinois, USA), Mayer Brown International LLP (England & Wales), Mayer Brown Hong Kong LLP (a Hong Kong limited liability partnership) and Tauil & Chequer Advogados (a Brazilian law partnership) (collectively, the “Mayer Brown Practices”). The Mayer Brown Practices are established in various jurisdictions and may be a legal person or a partnership. PK Wong & Nair LLC (“PKWN”) is the constituent Singapore law practice of our licensed joint law venture in Singapore, Mayer Brown PK Wong & Nair Pte. Ltd. Mayer Brown Hong Kong LLP operates in temporary association with Johnson Stokes & Master (“JSM”). More information about the individual Mayer Brown Practices, PKWN and the association between Mayer Brown Hong Kong LLP and JSM (including how information may be shared) can be found in the Legal Notices section of our website. “Mayer Brown” and the Mayer Brown logo are the trademarks of Mayer Brown. © 2025 Mayer Brown. All rights reserved.