MAYER | BROWN

RESPONDING TO AI INCIDENTS

AI & SECURITY SERIES

SPEAKERS

ANA BRUDER

PARTNER

MAYER BROWN LLP

MIKE DRISCOLL

SENIOR MANAGING DIRECTOR

FTI CONSULTING

STEPHEN LILLEY

PARTNER

MAYER BROWN LLP

AGENDA

1. Defining AI Incidents
2. Responding to AI Incidents
3. Preparing to Respond to AI Incidents

01

DEFINING AI INCIDENTS

KEY CATEGORIES OF AI INCIDENTS

Evasion Attacks

Prompt injection and other attacks intended to bypass AI security filters, safety guardrails, and developer instructions.

Poisoning Attacks

Attacks involving manipulation of training data or learning process to compromise a model.

Extraction Attacks

Theft of information from an AI system, either about the model itself or the training data.

Supply Chain Attacks

Exploitation of dependencies, tools, and third-party components relied upon by AI systems.

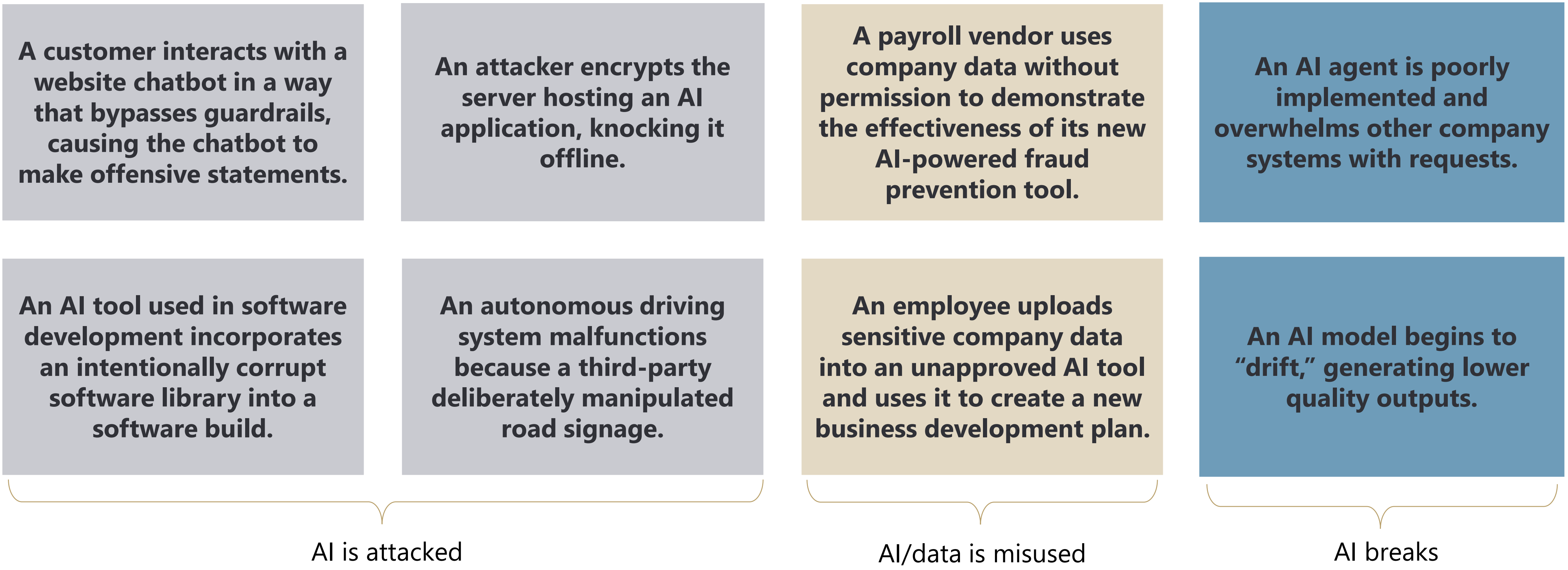
Abuse of Agentic AI

Misdirection, manipulation, or other exploitation of agentic tools within an organization.

Shadow AI

Unauthorized use of AI within an organization.

EXAMPLES OF AI INCIDENTS (OR ISSUES?)



A customer interacts with a website chatbot in a way that bypasses guardrails, causing the chatbot to make offensive statements.

An attacker encrypts the server hosting an AI application, knocking it offline.

An AI tool used in software development incorporates an intentionally corrupt software library into a software build.

An autonomous driving system malfunctions because a third-party deliberately manipulated road signage.

AI is attacked

A payroll vendor uses company data without permission to demonstrate the effectiveness of its new AI-powered fraud prevention tool.

An employee uploads sensitive company data into an unapproved AI tool and uses it to create a new business development plan.

AI/data is misused

An AI agent is poorly implemented and overwhelms other company systems with requests.

An AI model begins to “drift,” generating lower quality outputs.

AI breaks

DEFINING AN AI INCIDENT

At the time of writing, there are no widely accepted definitions of what constitutes an AI Incident, nor have any authoritative governmental bodies issued a definition. In contrast, cybersecurity incidents are well understood with key concepts such as the CIA Triad and authorities such as NIST providing guidance for how to think about threats and risks. GenAI security has clear overlaps with both ML security and cybersecurity; however, there are some divergences given the role of stochastic generation and natural language interaction with GenAI Applications.

- OWASP, GenAI Incident Response Guide (July 2025).

DEFINING AN AI INCIDENT

Cybersecurity Incident

"An occurrence that actually or imminently jeopardizes, without lawful authority, the confidentiality, integrity, or availability of information or an information system; or constitutes a violation or imminent threat of violation of law, security policies, security procedures, or acceptable use policies."

NIST SP 800-171r3

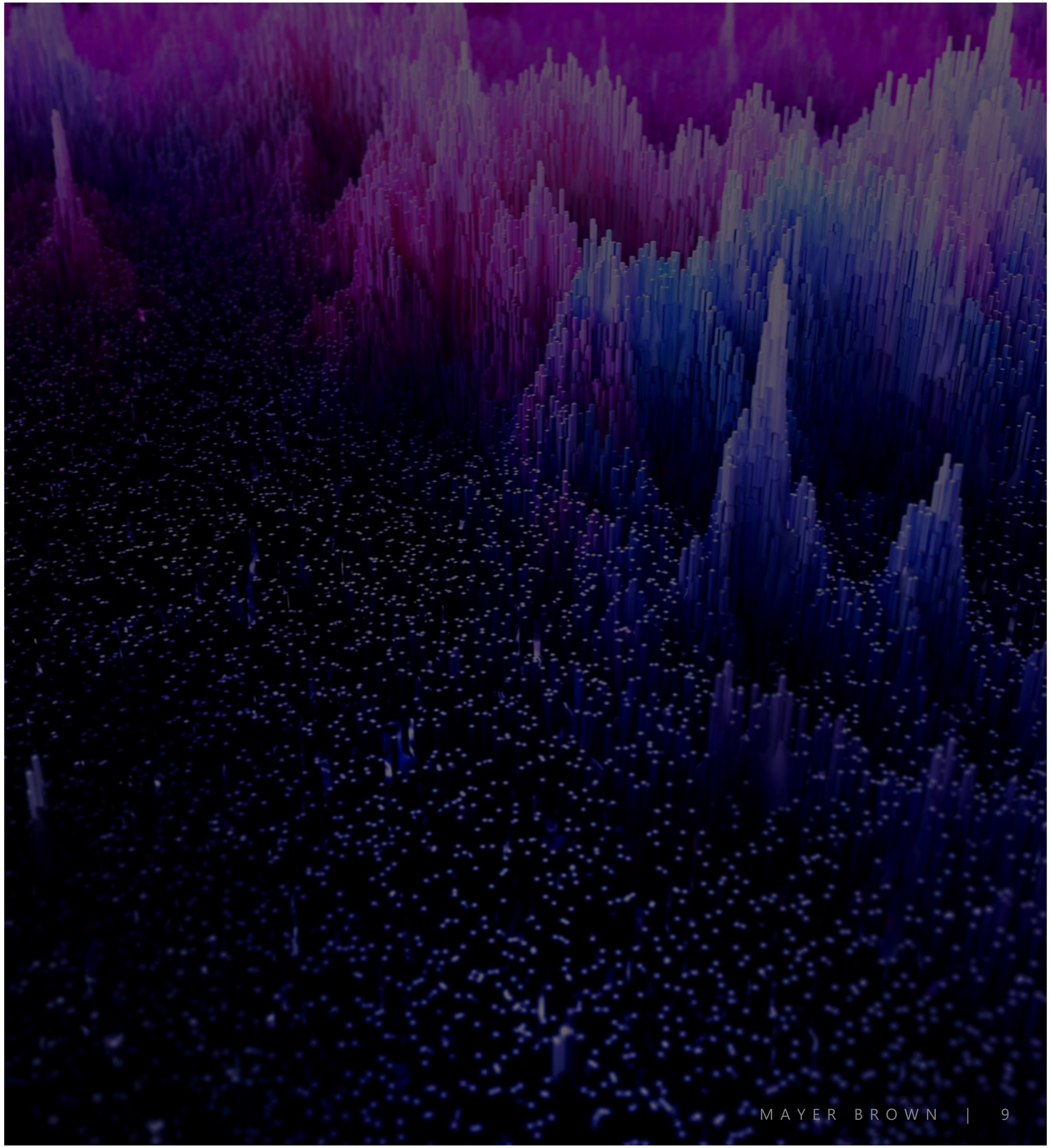
AI Incident

"**AI incident**" refers to an event, circumstance or series of events where the **development**, **use** or **malfunction** of AI system(s) directly or indirectly leads to **harm** to **individuals**, **property** or the **environment**, disruption of management or operation of **critical infrastructure**, violations of **human rights** or of laws to protect **fundamental rights, labor rights and IP rights**
OECD

A "**serious incident**" is an incident or **malfunctioning** of an AI system that directly or indirectly leads to **death** or serious harm to **health, property** or the **environment**; serious and irreversible disruption of management or operation of **critical infrastructure**, infringement of obligations under EU law intended to protect **fundamental rights**
EU AI Act

DEFINING AI INCIDENTS: WHY DOES IT MATTER?

- Companies need to be ready to respond to a wide range of issues, regardless of the labels applied.
- How issues are categorized nonetheless can have important practical consequences, including with respect to:
  - **Who will respond**—the roles and responsibilities of individual stakeholders such as security, legal, data governance, engineering in addressing the issue;
  - **How the company will respond**—which plan/process will be followed in resolving an issue, including the escalation paths that it will follow;
  - **What legal issues are presented**—whether regulatory or contractual obligations attach, including with respect to incident reporting; and whether the issue implies a policy violation for which action should be taken.



02

RESPONDING TO AI INCIDENTS

AI INCIDENTS: CHALLENGES TO ANTICIPATE

- **Breadth of potential incident types**
- **Uncertainty regarding roles and responsibilities**
- **Complexity of investigations**
  - Overall technical complexity of AI systems puts increased demands on forensic team
  - Limits on explainability may frustrate confirmation of root cause, etc.
  - Challenges reproducing an issue in a non-deterministic system
  - Reliance on third parties that operate AI tools procured by company
  - Challenges distinguishing between malicious activity and unexpected outcomes

AI INCIDENTS: CHALLENGES TO ANTICIPATE (CONT'D)

- **Limitations on containment and remediation**
  - It may be challenging to contain attacks without disabling functionality the business demands
  - Unlike a security vulnerability, fixing the issue that caused certain AI incidents, like prompt injection, does not necessarily avoid future problems
- **Communications challenges**
  - Difficulty explaining incident and response in light of technical complexity, limits on explainability, etc.
  - Challenges explaining facts and legal implications of incident to regulators who may not be experienced in dealing with these issues in the context of the laws they enforce
- **Internal risk multipliers:**
  - Rapid innovation may have created security debt/zombie applications with limited records or no current support
  - Internal pressures/incentives for rapid innovation may push against priorities in incident response

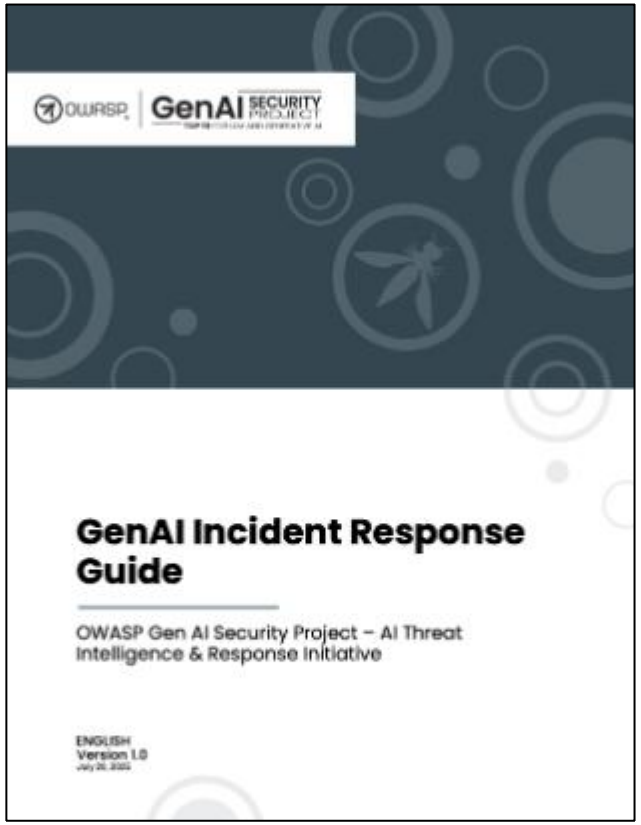
MEETING THE PARTICULAR CHALLENGES POSED BY AI INCIDENTS

- Response teams are likely to respond most effectively to AI incidents if they:
  - Calibrate and staff their response to the nature and severity of an AI incident;
  - Effectively balance predictability and rigor of response with flexibility to adapt processes considering novel demands posed by an incident;
  - Have ready access to appropriate technical expertise for forensic investigation, containment, remediation, etc.;
  - Manage applicable legal obligations while maintaining overall perspective on largest sources of legal risk.
- Strong cyber incident response capabilities are likely to be an important foundation for responding to these incidents—but undue reliance on these capabilities may backfire in some incidents.

RESOURCES FOR AI INCIDENT RESPONSE

CoSAI
COALITION FOR SECURE AI

AI Incident Response Framework, Version 1.0



OWASP | GenAI SECURITY PROJECT

GenAI Incident Response Guide

OWASP Gen AI Security Project – AI Threat Intelligence & Response Initiative

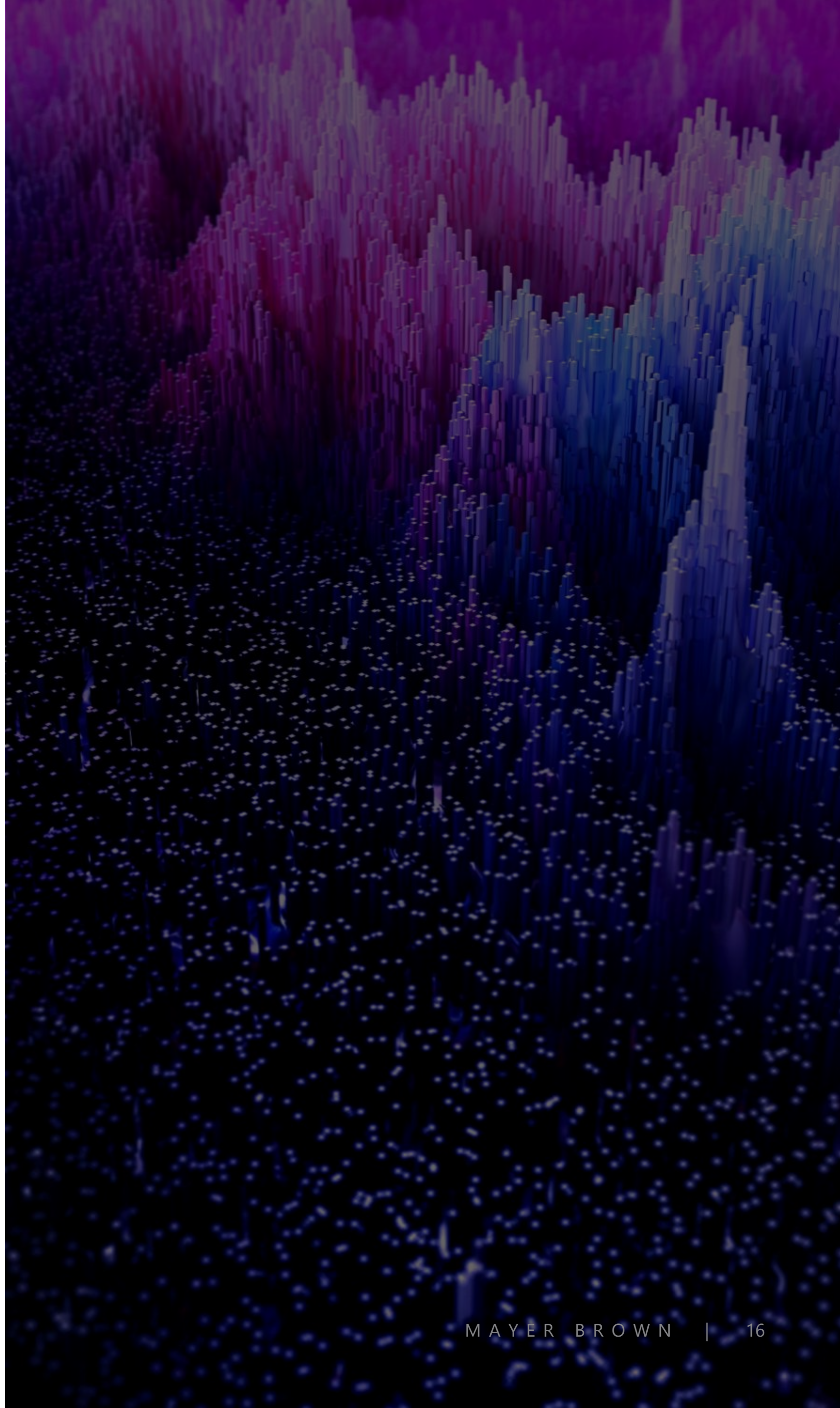
ENGLISH
Version 1.0
July 20, 2025

03

PREPARING TO RESPOND TO AI INCIDENTS

PREPARING TO RESPOND TO AI INCIDENTS

- Identify categories of AI incidents that the company may experience, associated AI systems, and risks to the company.
- Identify relevant stakeholders within the organization, assign applicable roles and responsibilities, and define escalation paths. Key groups may include:
  - AI teams: AI/ML engineers; data scientists
  - Information security
  - Information technology
  - Legal and compliance
  - Data owner / Relevant business unit (e.g., HR)
  - Communications
- Identify key legal obligations and risk-management priorities for response process.
- Update relevant processes and procedures, as well as operational playbooks.
- Ensure logging and other telemetry will support likely investigations.



PRACTICING RESPONSE TO AI INCIDENTS

- **Develop tabletop exercises and other scenario-based activities to allow key stakeholders to identify novel issues and challenges that may be posed by incidents involving AI systems.**
- **Key considerations for exercises:**
  - Develop realistic scenario that teases out key issues for company and addresses a priority risk area;
  - Incorporate correct stakeholders;
  - Capture and implement lessons learned—the exercise may serve as much to decide how you want to address novel issues as to practice existing skills.
- **Diversity of potential issues may counsel in favor of conducting multiple scenario-based sessions with smaller groups, in addition to a traditional enterprise-level tabletop.**

MAYER | BROWN