IAPP **AI Governance Global North America 2025**

Training 16-17 September

Workshops 17 September

Conference 18-19 September

BOSTON

Cybersecurity in the Age of AI

Ali Jessani
Shannon Togawa Mercer
Matthew F. Ferraro

Speakers

Ali Jessani
Counsel, WilmerHale
ali.jessani@wilmerhale.com

Shannon Togawa Mercer
Senior Counsel, OpenAI
stm@openai.com

Matthew F. Ferraro
Partner, Crowell & Moring
mferraro@crowell.com

Agenda

- **Impact of Generative AI**
  How AI is shaping the cyber landscape
- **Cyber Risks and Opportunities**
  Target, attack & defense categories
- **Regulation**
  U.S. and international perspectives
- **Risk Management**
  Frameworks for trustworthy AI
- **Incident Response**
  Preparing for when things go wrong

Impact of Generative AI on Cybersecurity

AI as Target

- Theft (model weights, models, data)
- Model inversion & privacy leaks
- Data poisoning
- Adversarial & backdoor attacks

AI as Attack Tool

- Generative phishing & scams
- Deepfake impersonations
- Polymorphic malware & WormGPT

AI for Defense

- Cybersecurity research
- Vulnerability & anomaly detection, patching
- Command line classification
- Detection and response
- PII detection and mitigation
- Predictive threat intelligence

[4] [5] [6]

AI as a Target

Theft

Attackers can: copy models (model extraction); exfiltrate / post model weights (in 2023, an LLM's model weights were posted to an internet forum); and use queries to recover sensitive training data (model inversion).

Data Poisoning

Attackers may inject malicious samples into training data to corrupt or bias models.

Adversarial & Backdoor Attacks

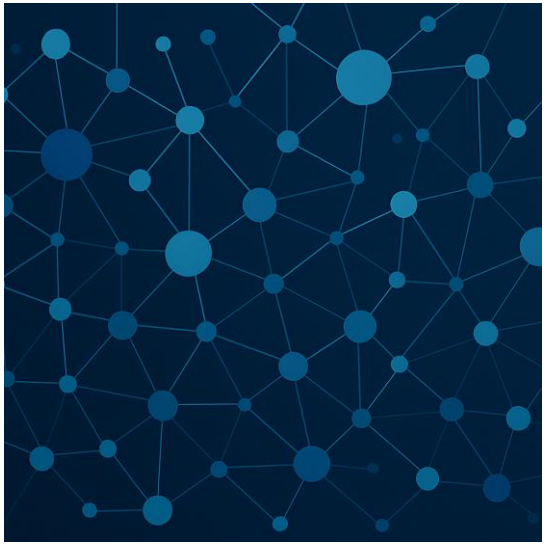
Tiny perturbations or hidden triggers can mislead models or open hidden backdoors during inference. For example, a model ignores a "Stop" sign.

AI as a Tool for Attacks

Phishing & deepfakes
Attackers craft convincing emails and audio/video deepfakes at scale, eroding trust and enabling social engineering.

Adaptive malware & WormGPT
Attackers leverage generative models to write polymorphic code, bypassing signature-based defenses.

Research and reconnaissance
Attackers using AI to analyze open-source data to tailor attacks, automating target selection and increasing success rates.



AI as a Tool for Defense

- Cybersecurity research
- Vulnerability & anomaly detection, patching
- Command line classification
- Detection and response
- PII detection and mitigation
- Predictive threat intelligence

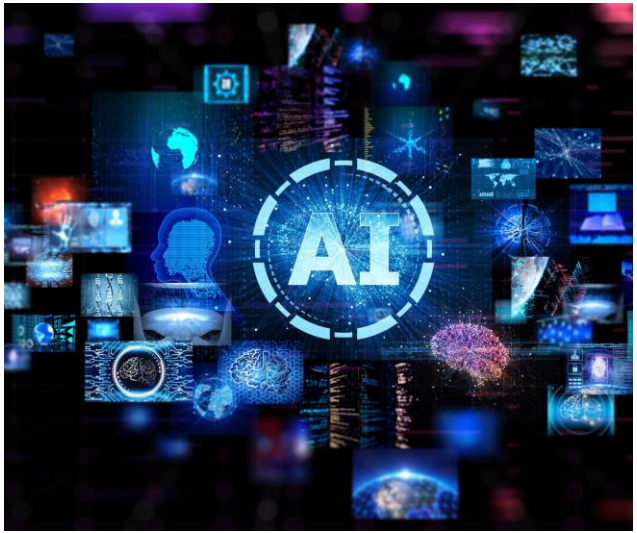
U.S. AI Regulation

Kaleidoscope of law-making and rulemaking
Federal legislature & Federal regulators; State legislatures & State regulators; International regulators; Plaintiffs' bar/Suits

Patchwork state landscape
A number of states (e.g., California, Texas, Arkansas, Colorado) are passing legislation regulating the use of AI, including (in some cases) specific security requirements.
The U.S. Senate did not seek to preempt state laws or to pass a "moratorium" last term, keeping the locus of AI lawmaking in the states.

Federal movements
For several years, federal agencies have used existing law to regulate AI and cybersecurity (including FTC Act and False Claims Act - Cyber Fraud Initiative).
The Trump administration passed the TAKE IT DOWN Act, rescinded prior executive orders, and released a 103-recommendation AI Action Plan focusing on innovation and infrastructure.

President Trump's AI Strategy Overview



- AI Action Plan and accompanying Executive Orders emphasize innovation, adoption, and competitiveness.
- The Plan calls for an examination of laws & regs that stifle AI but does not endorse a moratorium on / federal preemption of state AI laws and does not assert that the training of AI on copyrighted data constitutes "fair use."
- The Plan and EOs seek to fast-track the construction of energy and data center infrastructure and call both for greater exporting of the AI technology stack and enhanced controls on certain AI components.
- **Cybersecurity**: The Plan recommends ensuring AI used in homeland security be "secure by design"; creating an AI ISAC; issuing guidance on remediating AI-specific threats and coordinating on shared vulnerabilities; incorporating AI-specific scenarios into federal agency cybersecurity incident response plans.

President Biden's 2023 AI Executive Order

- Required developers of the most powerful AI systems to share their safety test results and other critical information with USG.
- NIST was directed to develop standards for AI systems.
- DOD and DHS were to pilot a plan to deploy AI to aid in the discovery and remediation of vulnerabilities in USG software.
- Commerce was to develop guidance for content authentication and watermarking of AI content to protect against fraud.

Cybersecurity Focused Initiatives

- Dual-use model developers had to report on cybersecurity protections.
- CISA worked with critical infrastructure to assess risks related to the use of AI in critical infrastructure, including cyber attacks.
- Sector-specific assessments by HHS and Treasury.
- Created AI Safety and Security Board at DHS.

International Regulations

EU AI Act

The EU AI Act's cybersecurity requirements mandate that high-risk AI systems must be secure, robust, and resilient against unauthorized attacks throughout their entire lifecycle. Key obligations include conducting thorough risk assessments, implementing technical safeguards against vulnerabilities like data poisoning or model evasion, ensuring systems can log operations for tracing breaches, and having fail-safe mechanisms

China

Have passed “Interim AI Measures”. Announced three specific cybersecurity standards for generative AI tools that will go into effect in November.

Other Jurisdictions

Brazil, Korea, Canada and others have adopted their own approaches to AI regulations, many of which include specific cybersecurity requirements.

Risk Management & Assessments

- **Map**
  Identify AI systems and contexts, catalogue models and data.
- **Measure**
  Assess risks: performance degradation, bias & privacy impacts.
- **Manage**
  Mitigate via policies, monitoring & controls.
- **Govern**
  Establish accountability, oversight & transparency across the lifecycle.

Offense vs Defense

Offensive AI

- Deepfake persuasion
- Dynamic malware & exploit generation
- Automated spear phishing
- Reconnaissance & social profiling

Defensive AI

- Real-time anomaly detection
- Threat hunting & correlation
- Automated SOC workflows
- Adversarial testing & red teaming

AI arms race

Both attackers and defenders iterate rapidly. As generative tools lower barriers to entry, defenders must leverage AI to maintain parity.

Incident Response

Why AI incidents are different

Non-deterministic outputs make it hard to distinguish malicious manipulations from benign anomalies. AI supply chains span multiple vendors, and telemetry is scarce, complicating detection.

- Prepare: define AI incidents, train teams, inventory assets
- Detect: monitor outputs & drift, run adversarial tests
- Contain: isolate compromised models, roll back versions
- Eradicate: patch prompts & integrations, retrain models
- Recover: restore functionality under monitoring
- Learn: document lessons and update policies

Questions to ask before an incident

Does my AI tool use or access “identifiable data”?

Many AI models can be trained on de-identified data, which reduces the potential risk in the event of a security incident (though there are still obligations related to “confidential” data, even if not identifiable information)

Who bears the responsibility?

Important to negotiate favorable security breach terms as part of the negotiation process (on both sides). Issues to pay attention to are breach notification costs, indemnification, and limitation of liability.

Have I taken steps to mitigate exposure?

Think about steps such as obtaining cyber insurance, proactively engaging vendors, and assigning roles and responsibilities in the event of an incident

Case Study: SalesLoft

- A major breach involving stolen authentication tokens from Salesloft's Drift chatbot impacted hundreds of services integrated with Salesloft, including Salesforce, Slack, Google Workspace, and cloud platforms like AWS and Azure.
- Google's Threat Intelligence Group reported the attackers, identified as UNC6395, began exfiltrating data from numerous corporate Salesforce instances starting August 8, 2025, prompting warnings to immediately invalidate all existing tokens.
- Salesloft products, used by over 5,000 customers, were blocked from integration with Salesforce, Slack, and Pardot as a protective measure, following the detection of stolen tokens and data theft.
- The incident is linked to a wave of social engineering and extortion attacks, with threat groups like ShinyHunters and Scattered Spider suspected, but exact attribution remains unconfirmed.
- Experts highlighted that attackers exploited legitimate user access tokens to move undetected between systems, emphasizing ongoing risks of "authorization sprawl".

Conclusion & Key Takeaways

- **Remain alert and flexible**
  Generative AI enhances both attack and defense strategies. Organizations need to constantly evolve by incorporating AI into their cybersecurity practices and surveillance.
- **Ensure compliance and foster collaboration**
  Stay abreast of changing regulations across different regions and partner with multidisciplinary teams to establish responsible AI governance frameworks.
- **Anticipate and prepare for challenges**
  Create incident response plans that consider AI-specific threats, conduct adversarial testing regularly, and continuously update knowledge as threat landscapes develop.

Questions?

How did things go?
(We really want to know)

Did you enjoy this session? Is there any way we could make it better? Let us know by filling out a speaker evaluation.

1. Open the IAPP Events app.
2. Select **IAPP AI Governance Global North America 2025**.
3. Tap "Schedule" on the bottom navigation bar.
4. Find this session. Click "Rate this Session" within the description.
5. Once you've answered all three questions, tap "Done".

Thank you!