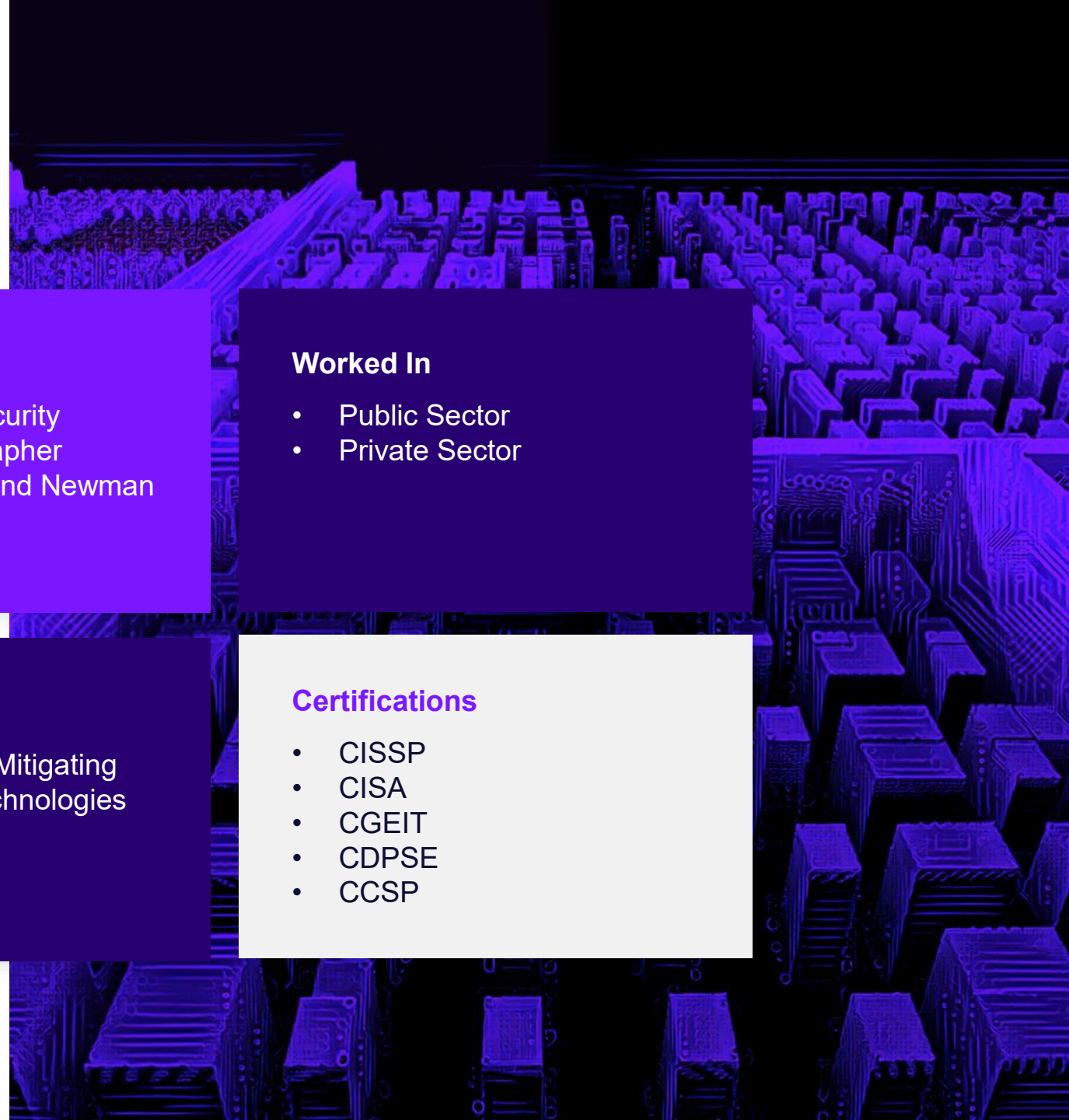
Keeping Your AI in Check:
No, It Shouldn’t Know Your Passwords

David Santeramo
Network and Security Leader

About Me



Experience

- Executive Risk Advisor
- Security Operations Leader
- Security Architect
- Network Administrator

30 Years

- Information Security
- Navy Cryptographer
- Bolt, Baranek and Newman

Worked In

- Public Sector
- Private Sector

20 Years

- Technical Leadership Roles at General Electric

Specialty

- Identifying and Mitigating Risk in New Technologies

Certifications

- CISSP
- CISA
- CGEIT
- CDPSE
- CCSP

Agenda

1. Risks

Phishing In The Age Of AI

Supply Chain – Compromised LLMs

Data Loss / Data Protection

User Access

Data Poisoning

State Actors Misuse

2. Remediation & Compliance

NIST AI Framework (Govern, Map, Measure, Manage)

Google SAIF

MITRE ATLAS

OWASP AI Top 10 Risks

EU AI Act

Actions You Can Take Today

Foundations of AI

01

What is AI?

What is AI?

The Discussion

- Present state of Risk surrounding AI
- What some of the risks are. From the training of models and data to threat actors
- Frameworks that are presently available to aid you and your organization in their adoption of AI

What Will Not Be Discussed

- Will not be delving into Gen AI. No productivity discussions around Grok, Gemini, ChatGPT, Copilot and Claude. We will be using the tools however
- Security tools using AI
- Will not be discussing hardware so don’t expect to hear a lot about chips and processing. Some discussion however regarding coolant.

AI stages of development

Artificial Intelligence

is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.

Machine Learning

Where a computer learns from data. Fraud detection, predictive analytics. Quality of data over quantity of data

Deep Learning

A flavor of Machine Learning that uses neural networks. Detects complex patterns and relationships

Generative AI

Creation of new content. Interative

Present Stake of Risk Surrounding AI?

Claude

Adversarial Attacks

Data Poisoning

Model Extraction & Theft

Prompt Injection

Privacy Attacks

INFRASTRUCTURE ATTACKS

Gemini

Algorithmic Bias & Discrimination

Misinformation & Manipulation (Deepfakes)

Data Quality & Integrity (Data Poisoning)

LACK OF ADEQUATE GOVERNANCE & OVERSITE

Copilot

Automated Cyberattacks

Deepfake & Social Engineering

Data Privacy Concerns

Bias & Model Poisoning

AI-Powered Malware

OVERRELIANCE ON AI

ChatGPT

Adversarial Attacks

Data Poisoning

Model Inversion & Data Leakage

Abuse of AI Tools

Supply Chain Vulnerabilities

INSUFFICIENT ROBUSTNESS & MISUSE

Summarize the cybersecurity risks of AI?

Same question. Different methods of training the model produced differing results.

The State of AI

Think about it…

Where was Cloud 10 years ago?

Where was the Internet 30 years ago?

Organizations are confronting AI risks today much like they tackled cloud adoption a decade ago—and the Internet three decades before that: **with caution, adaptation, and evolving governance.**

Risks

02

Phishing in the Age of AI

What we have been taught to look out for….

The Display Name

Salutations & Signatures

Typos

HTTPS or the Lock

Look Before Clicking

These are what we have been trained to look for… no longer

Phishing in the Age of AI

AI-Powered Phishing

Example: An attacker uses AI to scrape vast amounts of public information (social media, corporate websites, news articles) about a target. They then use a Large Language Model (LLM) like ChatGPT to craft an email perfectly mimicking the tone, style, and common phrases of a trusted colleague, CEO, or vendor. The email might reference recent events, projects, or even personal interests of the recipient, making it incredibly convincing.

Example: An AI could generate thousands of unique, personalized phishing emails simultaneously, each tailored to an individual's role, company, or even recent online activity, making mass phishing campaigns feel like highly targeted spear phishing. This automation significantly reduces the time and effort for attackers.

Deepfake Voice Phishing (Vishing)

Example: A finance employee receives a phone call or voicemail that sounds exactly like their CEO, requesting an urgent transfer of funds to a new, seemingly legitimate account. The AI can mimic the CEO's accent, cadence, and even emotional tone, making it extremely difficult to discern as fake.

Real-world Case: In 2019, an energy firm's CEO was tricked into transferring €220,000 after receiving a deepfake voice call from someone impersonating the CEO of the parent company, complete with a subtle German accent. Another case in 2024 involved a multinational firm losing $25 million due to a deepfake video call involving the CFO and other senior staff.

Supply Chain Risk | Compromised LLMs

What is LLMJacking?

Specific type of cyberattack that targets LLMs. It essentially refers to the hijacking or unauthorized use of cloud-based LLM resources.

Stolen Credentials

Attackers compromise cloud account credentials (like AWS keys, API keys for OpenAI, Azure, Google Cloud, etc.) often through traditional phishing, malware, misconfigurations, or by finding them exposed in public repositories (e.g., GitHub, Pastebin).

Exploiting Vulnerabilities

While less common for LLMjacking specifically (more common for prompt injection), general cloud security vulnerabilities in the infrastructure hosting the LLM can also be exploited.

Cost Avoidance

Running powerful LLMs is expensive, with costs potentially reaching tens or even hundreds of thousands of dollars per day for heavy usage. Attackers use the victim's resources to run their own LLM applications or queries, passing the hefty bill to the unsuspecting victim.

Developing or Optimizing Malicious Tools

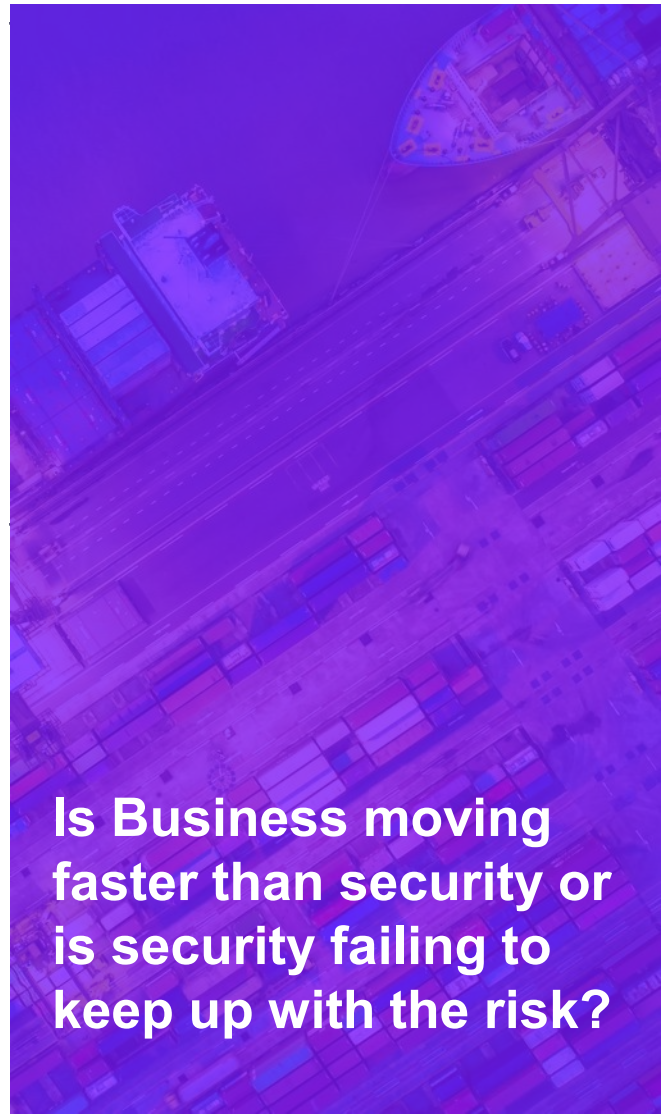
Using the LLM to write or refine malware, develop exploit code, or create sophisticated phishing scripts.

Data Poisoning

If the attacker gains sufficient privileges, they might attempt to inject incorrect or biased data into the LLM's training or fine-tuning process, subtly altering its future behavior or outputs for their benefit.

Stealing Sensitive Information

If the compromised credentials or LLM has access to internal data, attackers could try to extract sensitive or proprietary information.



Data Loss Prevention in an AI World

Language-Based Data Leaks

AI models can unintentionally expose sensitive information through summarization, paraphrasing, or translation.

Shadow AI Risks

Employees using unauthorized AI tools may inadvertently share confidential data.

AI-Powered Cyber Threats

Attackers leverage AI to bypass traditional security measures, making data breaches more sophisticated.

Simple question – How many of you know where all of your personal information is located?
Passports, SSN cards, birth certificates….

Data Loss Prevention (DLP) refers to a set of security measures designed to detect, prevent, and mitigate the unauthorized access, transmission, or leakage of sensitive data. It helps organizations safeguard confidential information from cyber threats, insider risks, and accidental exposure.

Accessing Data | Human & Non-Human | API Access

Synthetic Identity Fraud

AI speeds up the creation of convincing, entirely fake identities by combining real and fictional personal data, including generating authentic-looking ID cards and utility bills.

Deepfakes

AI generates highly realistic fake images, videos, and audio, making it easier to impersonate individuals and bypass traditional verification systems. This has led to a **704% surge in deepfake-driven "face swaps"** used to bypass identity verification in 2023.

Document Forgery

Generative AI tools facilitate the rapid and scalable alteration of digital documents like passports and utility bills. Digital forgeries now account for over **57% of all document fraud**, a **244% annual increase**.

How many of your organizations have reassessed your identity controls as you increasingly adopt AI?

How many still are tackling dormant identities?

How many of you know what a Non-Human Identity pertains to?

New ways to look at Identity Risk.
New tactics and methods that are being used to breach Identity.

Data Poisoning

Data poisoning is a type of adversarial attack in which malicious actors deliberately introduce corrupted or deceptive data into an AI model's training dataset, compromising its accuracy, reliability, or behavior.

SCENARIO

An attacker repeatedly submits emails containing legitimate words or phrases but labels them as "spam."

EFFECT

The spam filter, when retrained on this poisoned data, starts to misclassify legitimate emails as spam, or conversely, allows more actual spam to pass through, reducing its overall effectiveness. This was seen in attempts to poison Google's Gmail spam filter

SCENARIO

Subtle, almost imperceptible changes (e.g., adding a few pixels) are made to images of stop signs in the training data, but these poisoned images are labeled as "speed limit 45."

EFFECT

When the autonomous vehicle encounters a real stop sign with a similar subtle alteration (e.g., a sticker on the sign), it might misinterpret it as a speed limit sign, potentially leading to dangerous acceleration instead of stopping. This is a "backdoor" attack, where a specific "trigger" causes a desired misclassification.

State of Sponsored Risks

Islamic Republic of Iran

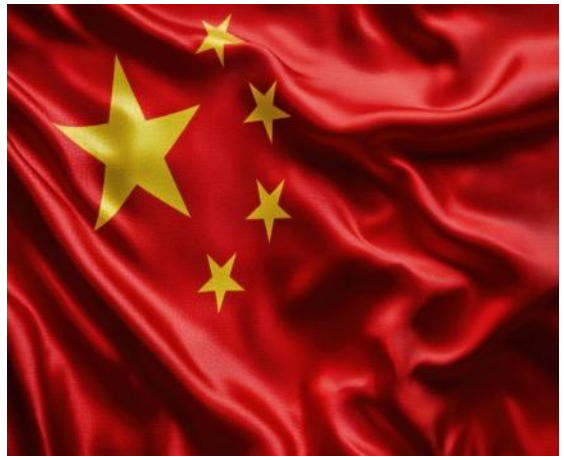
Coding and scripting (python, PowerShell)

Vulnerability research

Research about organizations (cybersecurity companies)

Research about warfare defenses

Generating content (translations into Farsi, Hebrew, and English)



People's Republic of China

Reconnaissance into US industries

Generating code specific to compromise known vulnerabilities

Using AI to specifically target email

More than 24 known groups actively using Gen AI platforms

Deeper system access and post-compromise actions

Democratic People's Republic of Korea

Phishing and account compromise techniques

Vulnerability research on cryptocurrency and OT networks

Automation and scripting to acquire data from compromised accounts

Clandestine employee services (LinkedIn)

Nuclear technology and power plants in South Korea

Russian Federation

Rewriting malware into other languages

Payload crafting (encryption)

Cryptocurrency and Financially motivated

manipulated video and voice content in business email compromise (BEC) activities

Compromised LLMs

Remediation & Compliance

03

Compliance In A Constantly Shifting Landscape…

FEDERAL

Consumer Protection Laws
(Federal Trade Commission - FTC)

Anti-discrimination Laws
(Equal Employment Opportunity Commission - EEOC, Department of Justice - DOJ)

Health Insurance Portability and Accountability Act (HIPAA)

Gramm-Leach-Bliley Act (GLBA)

Children's Online Privacy Protection Act (COPPA)

Electronic Communications Privacy Act (ECPA)

NEW YORK STATE

New York AI Consumer Protection Act (A768/S1962): Prohibits AI algorithms from discriminating against protected classes.

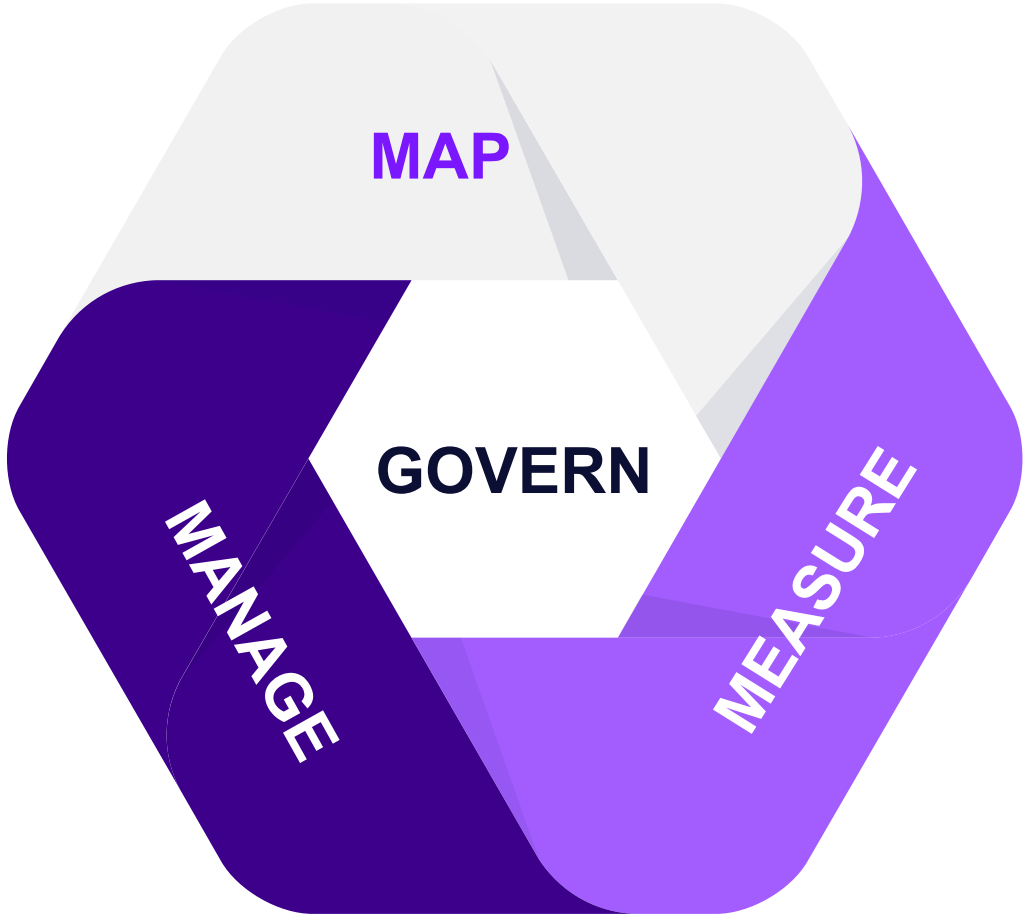
Senate Bill S1169: Regulates high-risk AI systems, mandates independent audits, and establishes enforcement mechanisms.

New York State Agency Oversight: Requires New York government agencies to review and report how they use AI software.

Emerging Federal AI-Specific Directives and Proposed Legislation

Executive Order on Removing Barriers to American Leadership in Artificial Intelligence (January 2025) • Generative AI Copyright Disclosure Act (Proposed) • American Privacy Rights Act (APRA) (Proposed) • TAKE IT DOWN Act (April 2025) • CREATE AI ACT

The NIST AI Risk Framework



First Draft published in January 2023

MAP

Context is recognized and risks related to the context are identified

MEASURE

Identified risks are assessed, analyzed or tracked

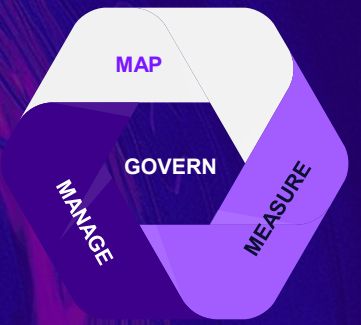
MANAGE

Risks are prioritized and acted upon based on a projected impact

GOVERN

A culture of risk management is cultivated and encouraged

The NIST AI Risk Framework



MAP

Understand and verify context of use for AI Systems

Identify system limits, beneficial applications and AI limitations

Address real world constraints and assess impacts of AI use

MEASURE

Employ diverse tools for AI risk analysis and ensure trustworthiness

Apply rigorous unbiased assessments and address risk

Establish TEVV processes and transparency

Continuously adapt to emerging risks and make informed decisions

MANAGE

Assess AI Risks, prioritize management and evaluate system viability

Document residual risk and strategize to minimize impact

Allocate resources, consider alternatives and manage third party risks

Monitor and improve AI risk treatments with effective communications with stakeholders

GOVERN

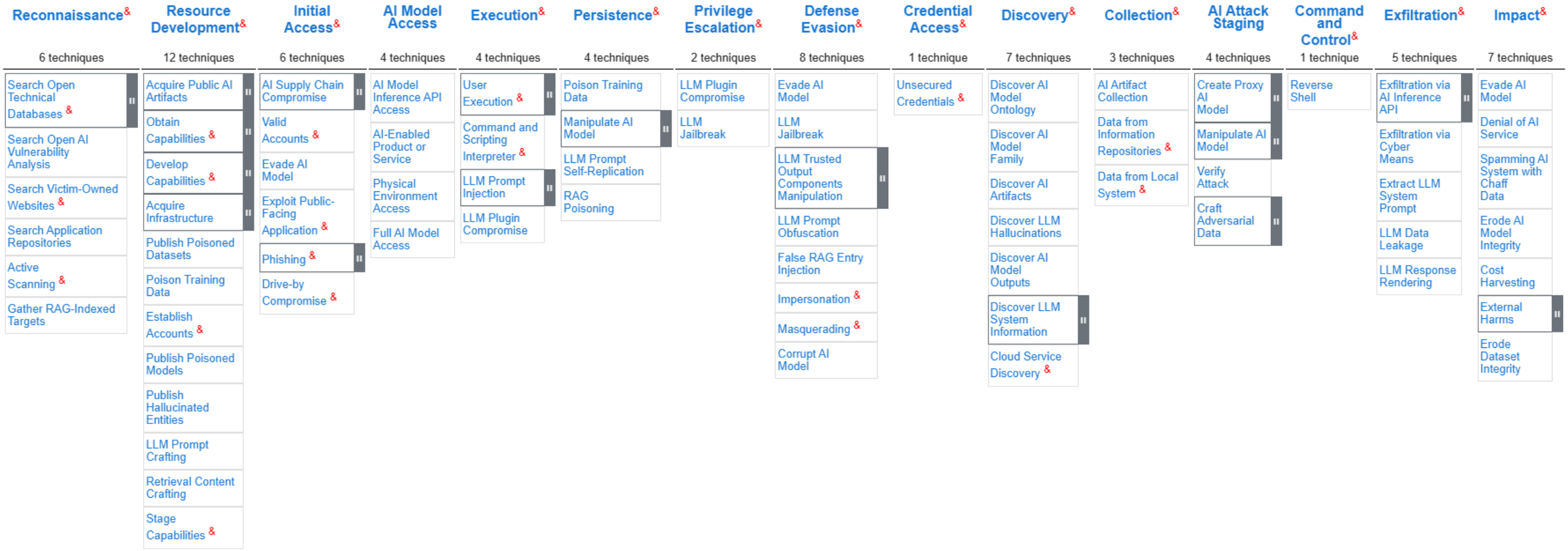
Cultivate a proactive risk management

Develop comprehensive processes to anticipate AI risks

Evaluate potential impacts and align with organizational values

Integrate technical design with principles and consider full lifecycle

MITRE ATLAS
(Adversarial Threat Landscape for Artificial-Intelligence Systems)

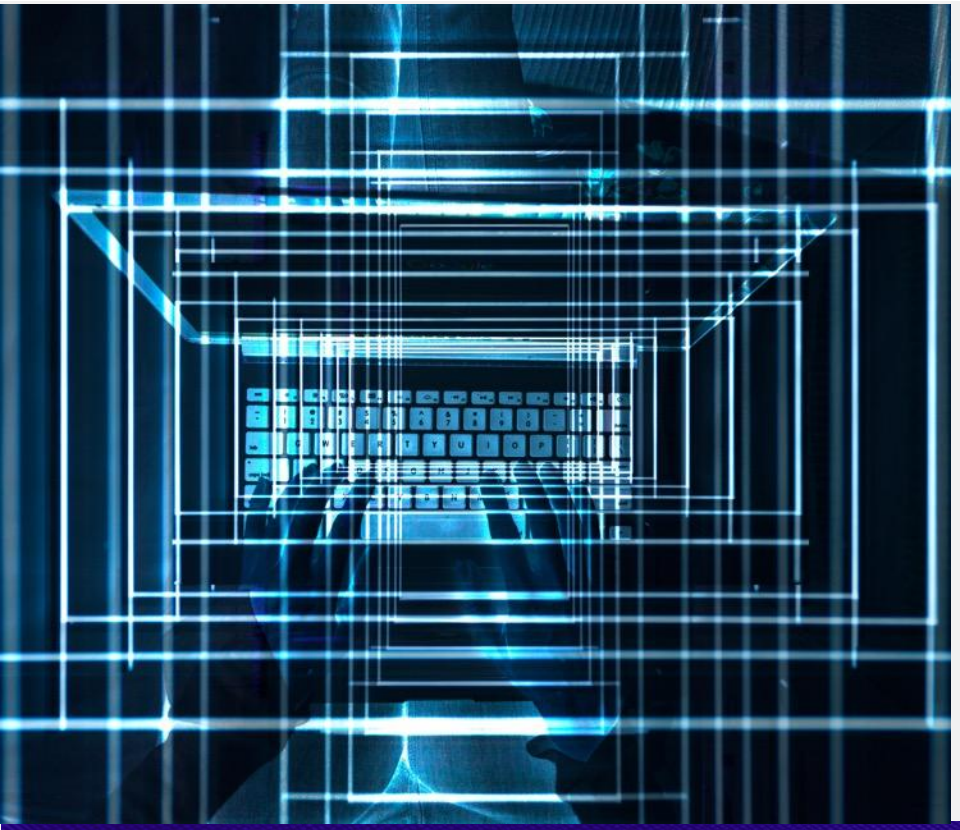


Initially released in 2021, ATLAS is designed to identify, classify, and mitigate adversarial threats to AI and ML systems

OWASP Top 10 AI Risks for 2025

ATTACK TECHNIQUE	MITIGATION
Sensitive Information	Encryption, authentication, Output Controls, DLP, Employee Training
Insecure Plug Design	Authentication, Rate Limiting, Least Privilege, Network Segmentation
Excessive Agency	Human in the Loop Design, Operational Boundaries, Logging
Overreliance	Confidence scoring, AI literacy training, Performance Metrics
Model Theft	Access Control, Least Privilege, Regular audits, AI gateways
Prompt Injection	Input filtering, length restrictions, Robust pre-training data, confidence scoring
Insecure Handling	Access Control, encryption, Secure coding practices, Version control
Model DOS	Load Balancing, auto scaling, input validation, DOS mitigation services
Data Poisoning	Outlier Detection, Anomaly Detection, Input validation
Supply Chain	Trusted Data Sources, Trusted libraries, code signing, secure APIs

Google’s Security AI Framework (SAIF)



Six Core Security Elements

SAIF focuses on:

- Expanding strong security foundations for AI systems.
- Extending detection and response to AI-related threats.
- Automating defenses to counter evolving AI risks.
- Harmonizing security controls across AI platforms.
- Adapting security measures for AI deployment.
- Contextualizing AI risks within business processes.

Industry Collaboration

Google has partnered with organizations like **Microsoft**, **OpenAI**, **IBM**, and **NVIDIA** to advance AI security standards.

Government & Policy Engagement

SAIF aligns with regulatory frameworks and contributes to AI security discussions with policymakers.

EU AI Act

The EU AI Act was provisionally agreed upon in December 2023 and is expected to be formally adopted and gradually phased in over the next few years, with different provisions coming into effect at different times.

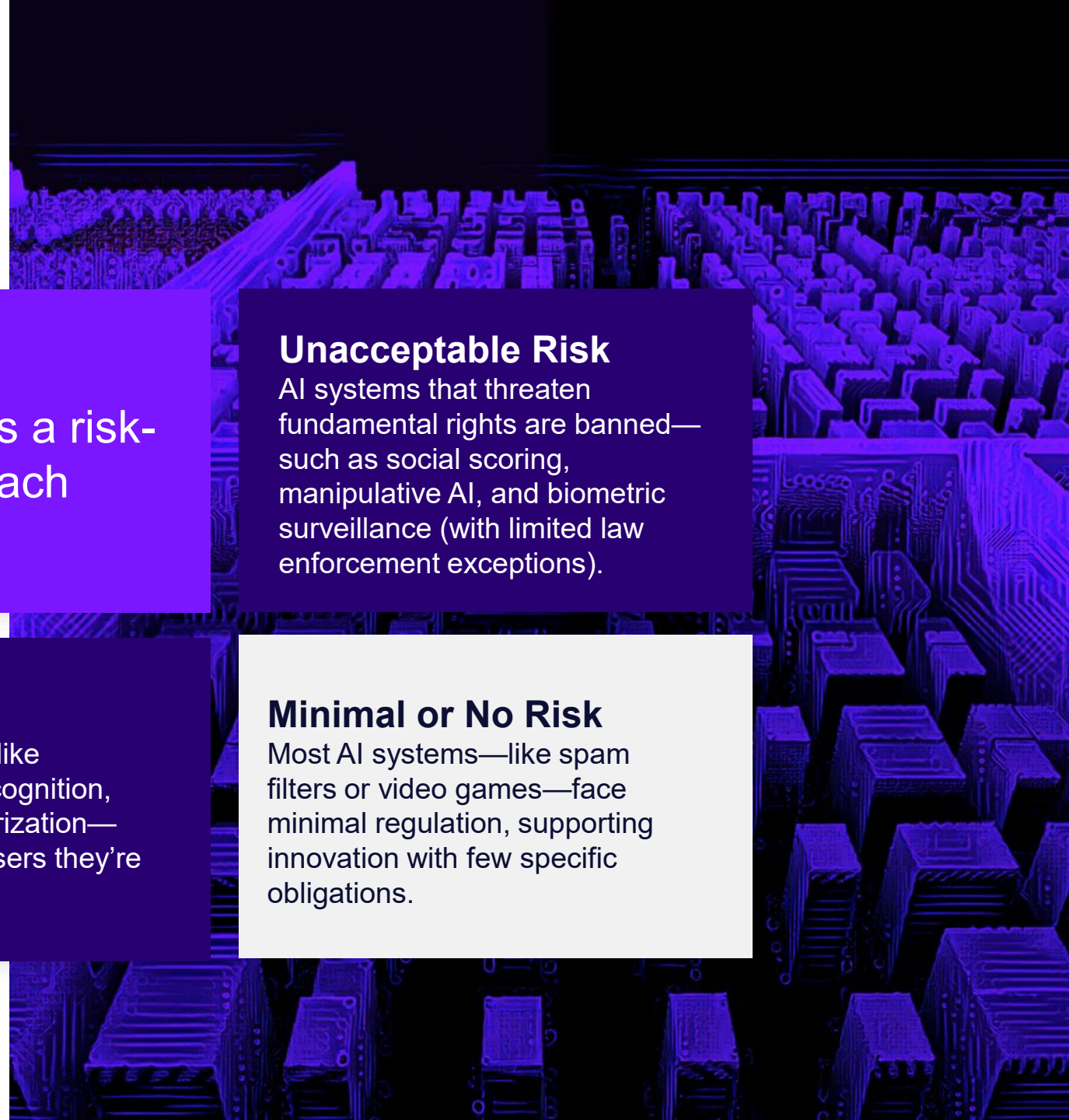
The Act takes a risk-based approach

Unacceptable Risk
AI systems that threaten fundamental rights are banned—such as social scoring, manipulative AI, and biometric surveillance (with limited law enforcement exceptions).

High-Risk
High-risk AI systems that may impact safety or fundamental rights face strict requirements—covering areas like infrastructure, education, employment, public services, law enforcement, and democratic processes.

Limited Risk
Certain AI systems—like chatbots, emotion recognition, and biometric categorization—must clearly inform users they're interacting with AI.

Minimal or No Risk
Most AI systems—like spam filters or video games—face minimal regulation, supporting innovation with few specific obligations.



AI Privacy Laws & Regulations

3 Elements

1. Right To Access
2. Right To Correct
3. Right To Object

Under GDPR, organizations must have a lawful basis for processing data (consent, contractual necessity or legal obligation) AND they must also implement appropriate measure to protect personal data

GDPR

New international standard that provides a framework for organizations to establish, implement, maintain, and continually improve an **Artificial Intelligence Management System** (AIMS)

Applicable to all organizations. No matter the size

Risk Management: Helps organizations identify and mitigate risks associated with AI

Efficiency & Compliance: Provides a structured approach to managing AI systems while ensuring compliance with regulations

ISO/IEC 42001

Applies to for profit enterprises, operating in the state, having **annual gross revenues over $25 million** OR **earning more than half of their revenue** from selling consumer data.

Data of state residents that are being incorporated into LLMs without consent.

California Consumer Privacy Act 2020

CCPA

Enacted in 2021 in China

Aims to uphold rights to personal information, standardize data handling procedures

Addresses the processing of personal data inside the borders AND, in limited cases, extends outside.

Imposes responsibilities on data handlers (security measures, auditing)

Personal Information Protection Law 2021

PIPL

Preparing Your Organization for AI Compliance

Comprehensive Risk Assessment

Identify potential risks associated with AI systems, including ethical concerns, security vulnerabilities, and compliance challenges.

AI Incident Response

Develop response plans and perform tabletop exercises for AI related incidents inside your organization.

AI Governance Policies

Define clear policies for AI development, deployment, and monitoring to align with regulatory requirements and ethical standards.

Know Your Data

How is it Accessed

Understand where all of your sensitive data is located and how it is accessed.

Integrate AI Management into Business Processes

Ensure AI governance is embedded within existing organizational frameworks to support responsible AI use.

Continuous Monitoring & Improvement

Regularly assess AI systems for biases, security threats, and performance issues, making necessary adjustments. (Data Protection)

Train Employees on AI Compliance

Educate staff on AI ethics, security best practices, and regulatory obligations to foster a culture of responsible AI use.

Thank You!

pellera