AI & CYBERSECURITY
THE INTERSECTION OF DEFENSE, THREATS, AND PROTECTION

Razi Rais

About

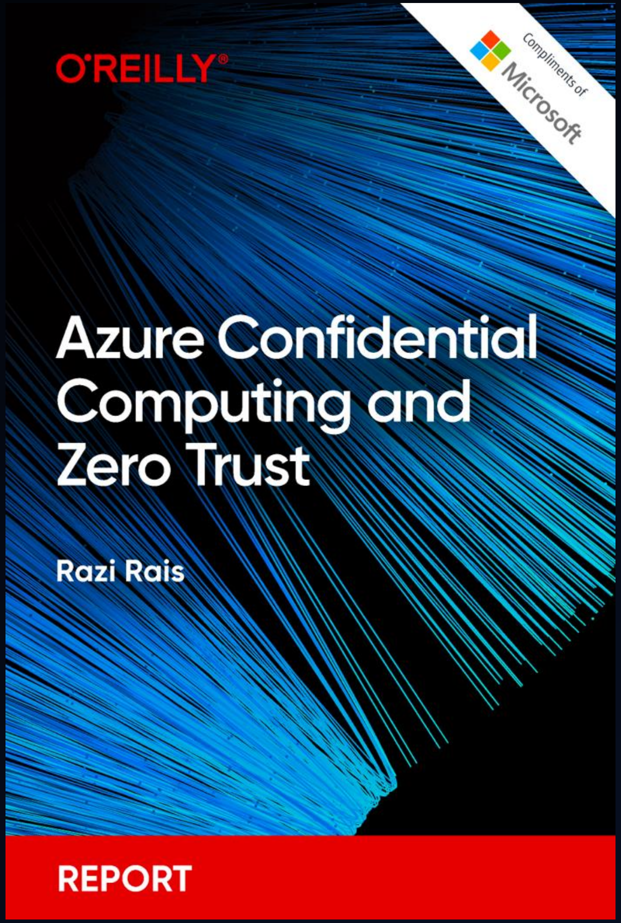
❖ **20+ years of experience** in the software development, architecture, and product management.

❖ **10+ years at Microsoft** in various teams as a software engineer, product manager, and architect.

❖ Currently working as a senior technical product manager at Microsoft helping businesses secure their digital identities at cloud scale.

❖ **Published author, speaker, and trainer**.

Cybersecurity books



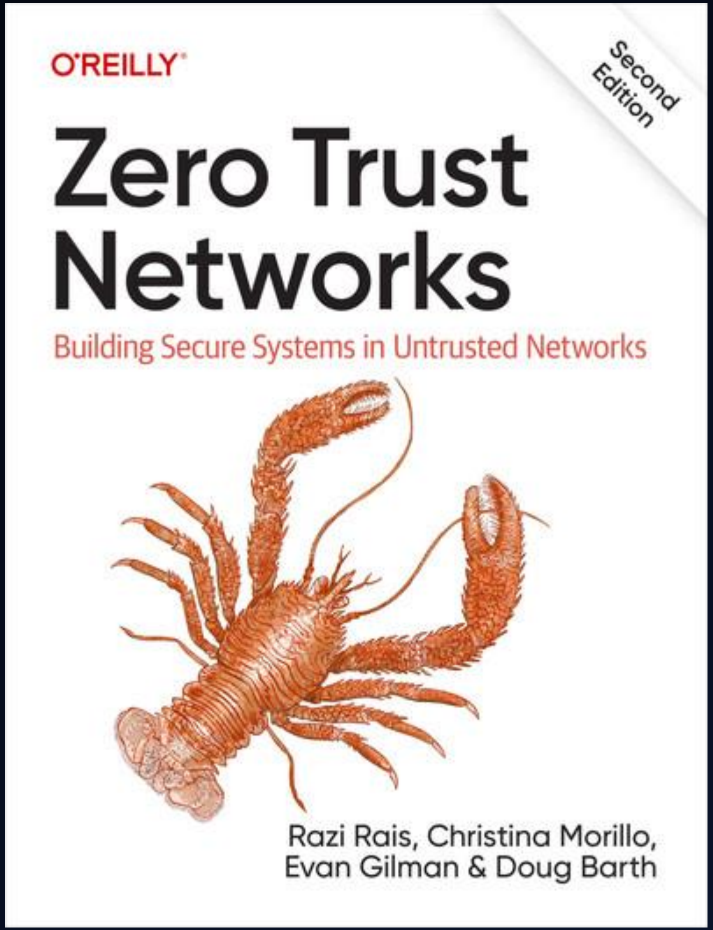
O'REILLY®

Compliments of Microsoft

Azure Confidential Computing and Zero Trust

Razi Rais

REPORT



O'REILLY®

Second Edition

Zero Trust Networks

Building Secure Systems in Untrusted Networks

Razi Rais, Christina Morillo, Evan Gilman & Doug Barth

Read Online

Read Online

⚡ Let's Connect!

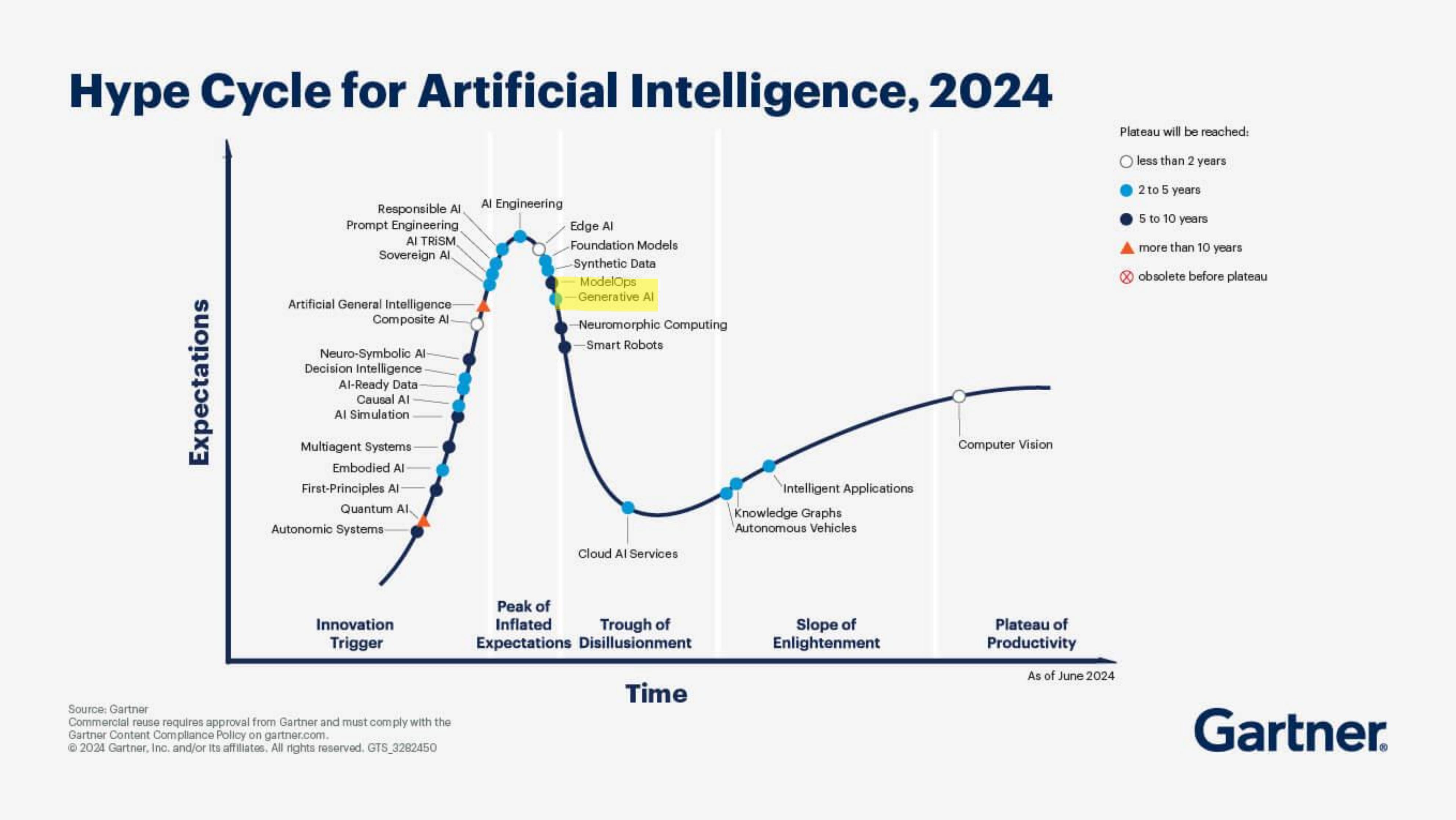
Looking for mentorship in AI and cybersecurity? Need an expert speaker for your next event? Working on an AI-powered cybersecurity project and need strategic guidance? Let's collaborate—reach out today on LinkedIn!

https://www.linkedin.com/in/razirais/

Agenda

- ✓ Role of AI in Cybersecurity: Security of AI + AI in Security
- ✓ Learning Resources
- ✓ Discussion

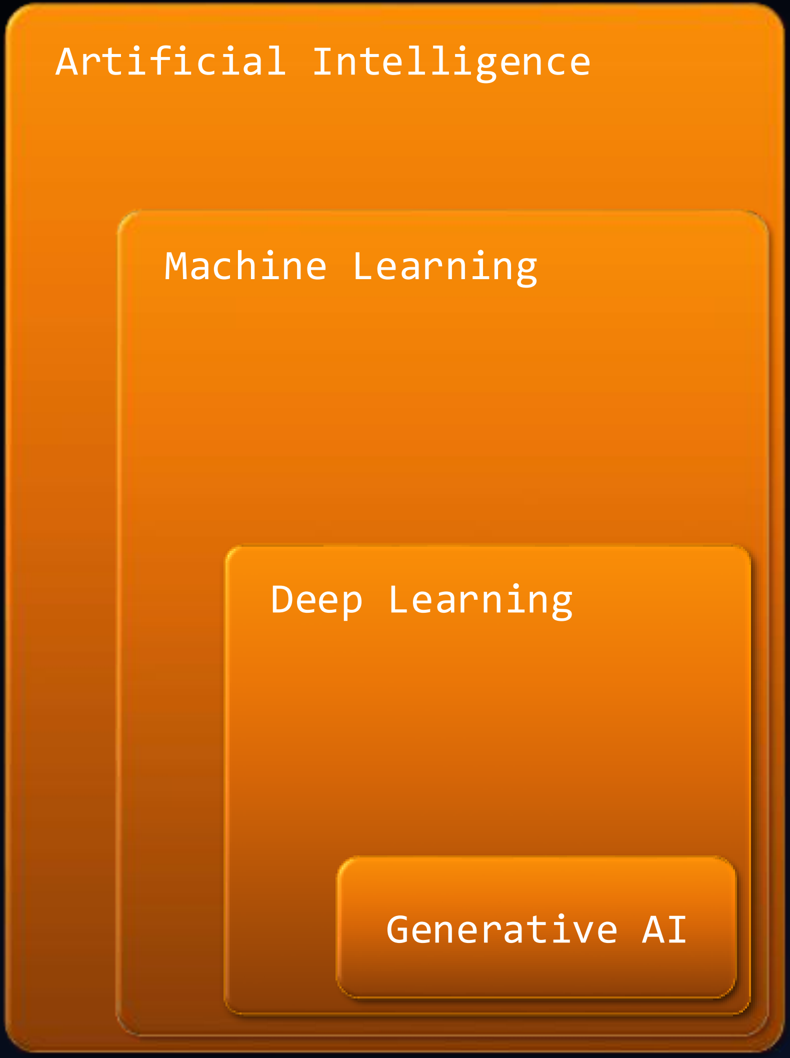
Gartner – Hyper Cycle for AI 2024



Source: Gartner
Commercial reuse requires approval from Gartner and must comply with the Gartner Content Compliance Policy on gartner.com.

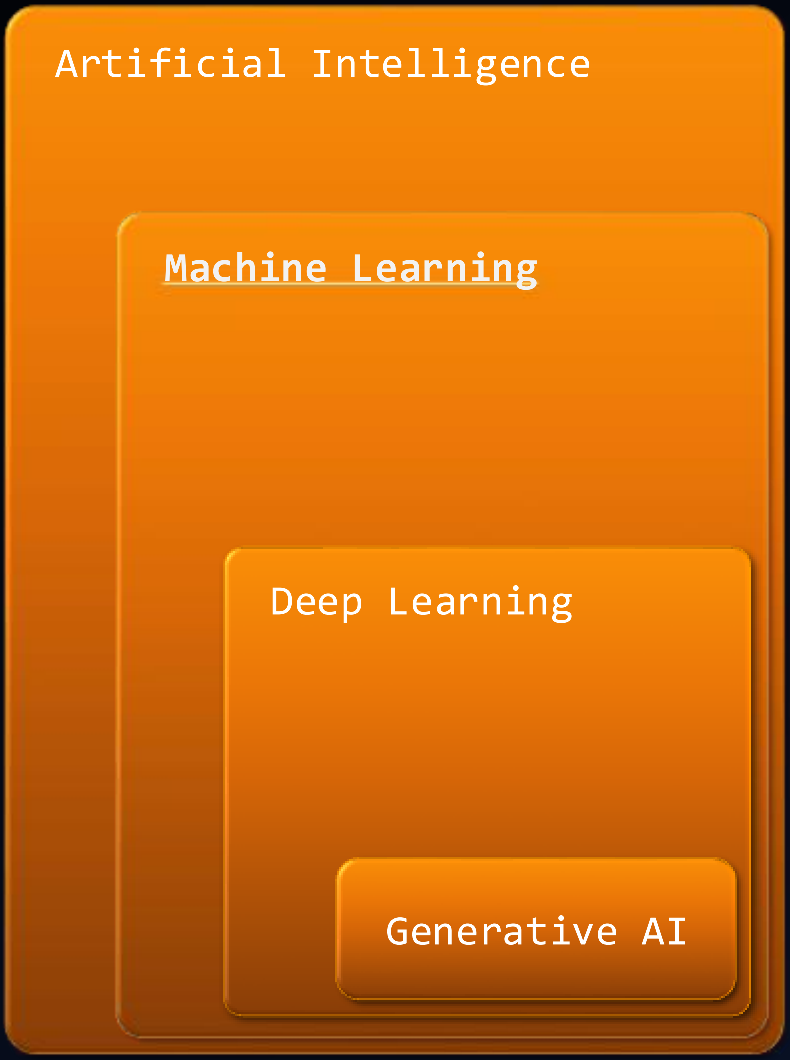
https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

Artificial Intelligence (AI)



- **Artificial Intelligence**: The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence
- **Machine Learning**: Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions
- **Deep Learning**: A machine learning technique in which layers of neural networks are used to process data and make decisions
- **Generative AI:** Create new written, visual, and auditory content given prompts or existing data

Artificial Intelligence (AI) Cont.



Artificial Intelligence

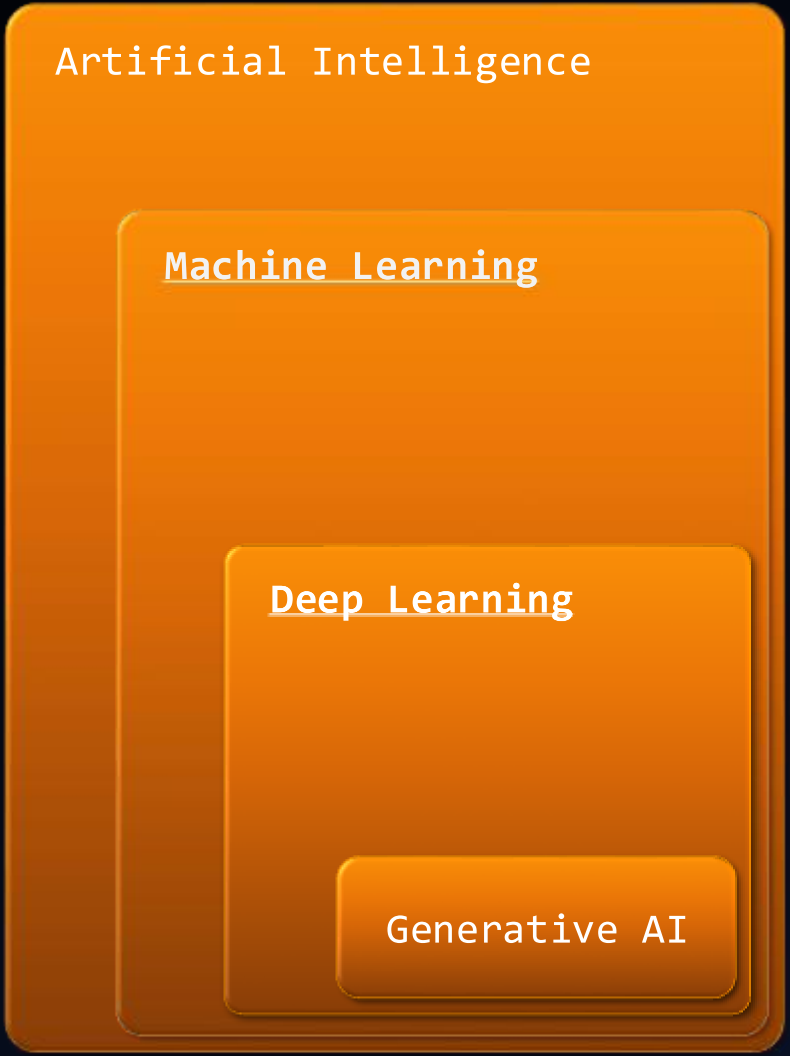
Machine Learning

Deep Learning

Generative AI

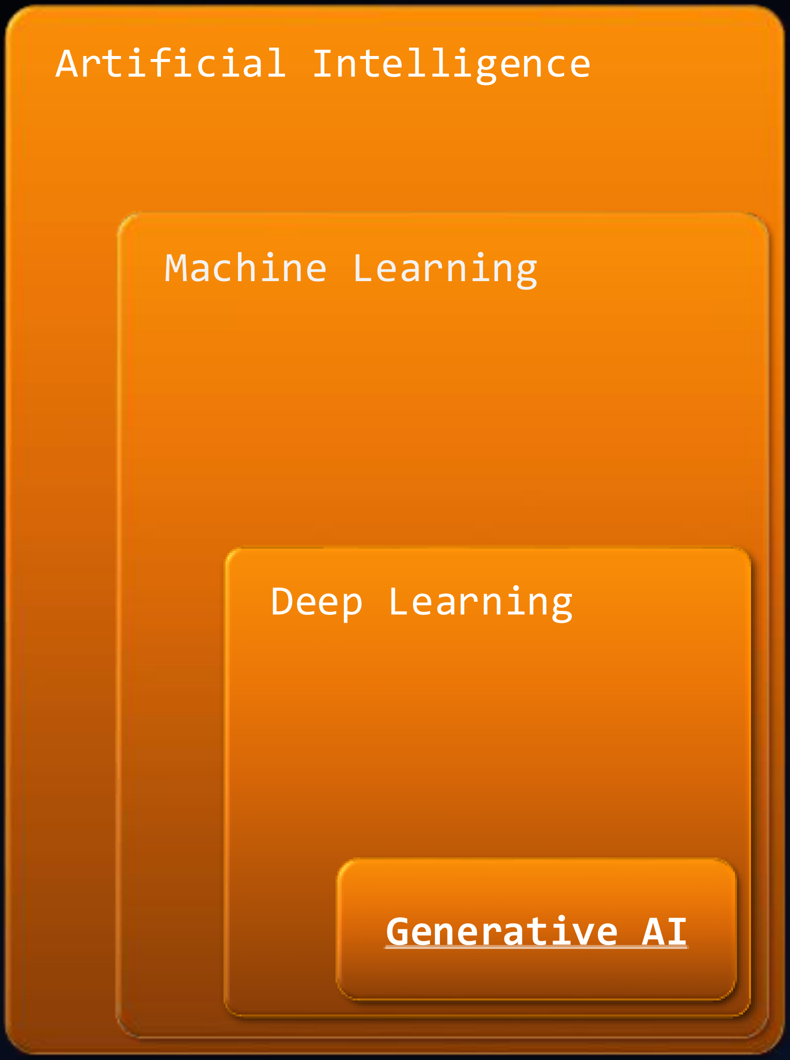
- **Machine Learning**: The first decade of the 2000s marked the rapid advance of various machine learning techniques that could analyze massive amounts of online data to draw conclusions – or “learn” – from the results. Since then, companies have viewed machine learning as an incredibly powerful field of AI for analyzing data, finding patterns, generating insights, making predictions and automating tasks at a pace and on a scale that was previously impossible.

Artificial Intelligence (AI) Cont.



- **Deep Learning:** The 2010s produced advances in AI's perception capabilities in the field of machine learning called deep learning. Breakthroughs in deep learning enable the computer vision that search engines and self-driving cars use to classify and detect objects, as well as the voice recognition that allows popular AI speech assistants to respond to users in a natural way.

Artificial Intelligence (AI) Cont.



Artificial Intelligence

Machine Learning

Deep Learning

Generative AI

- **Generative AI:** Building on exponential increases in the size and capabilities of deep learning models, the 2020s will be about language mastery. The GPT-4 language model, developed by OpenAI, marks the beginning of a new phase in the abilities of language-based AI applications. Models such as this will have far-reaching consequences for business, since language permeates everything, an organization does day to day—its institutional knowledge, communication and processes.

Terminology

- Model → Patterns or relationships in data
- Large Language Models (LLM)
  - Large: More data than can be manually labeled
  - Language: Match context and words (e.g., word prediction, creative writing)
  - Model: Semi-supervised learning

ChatGPT

- ChatGPT: AI Chatbot, developed by OpenAI, trained to perform conversational tasks and creative tasks
- Conversation-in and message-out
- Trained over 175 billion machine learning parameters
- GPT-4 and above are multimodal (e.g., images + text)

Domain Specific LLMs

- While generic LLMs (like ChatGPT) are great fit for general queries but they cannot understand the specific context beyond the massive datasets they are trained with.
- If you didn’t train a language model with domain specific data, the results may be less than ideal.
- So, we need custom models with a better language understanding of a specific domain (Finance, Healthcare, Cybersecurity etc.)

Domain Specific LLMs (Cont.)

- BlackrockGPT
- GoldmanSachsGPT
- StripeGPT
- MorningstarGPT
- RobinhoodGPT
- VanguardGPT
- SoFiGPT
- ChubbGPT
- RevolutGPT
- ChatLAW
- KAI-GPT
- FinGPT
- ClimateBERT

AI in Security

Using AI to enhance cybersecurity, such as preventing cyberattacks, optimizing security processes, and improving security resilience.

- Intrusion Detection and Prevention
- User Behavior Analytics
- Vulnerability Assessment
- Cyber threat intelligence
- Phishing Protection

SOC Use case: Security posture management

- Assist with evaluating whether an organization is vulnerable to known vulnerabilities and exploits.
- Assist in risk prioritization
- Assist in resolving weaknesses by making specific advice.

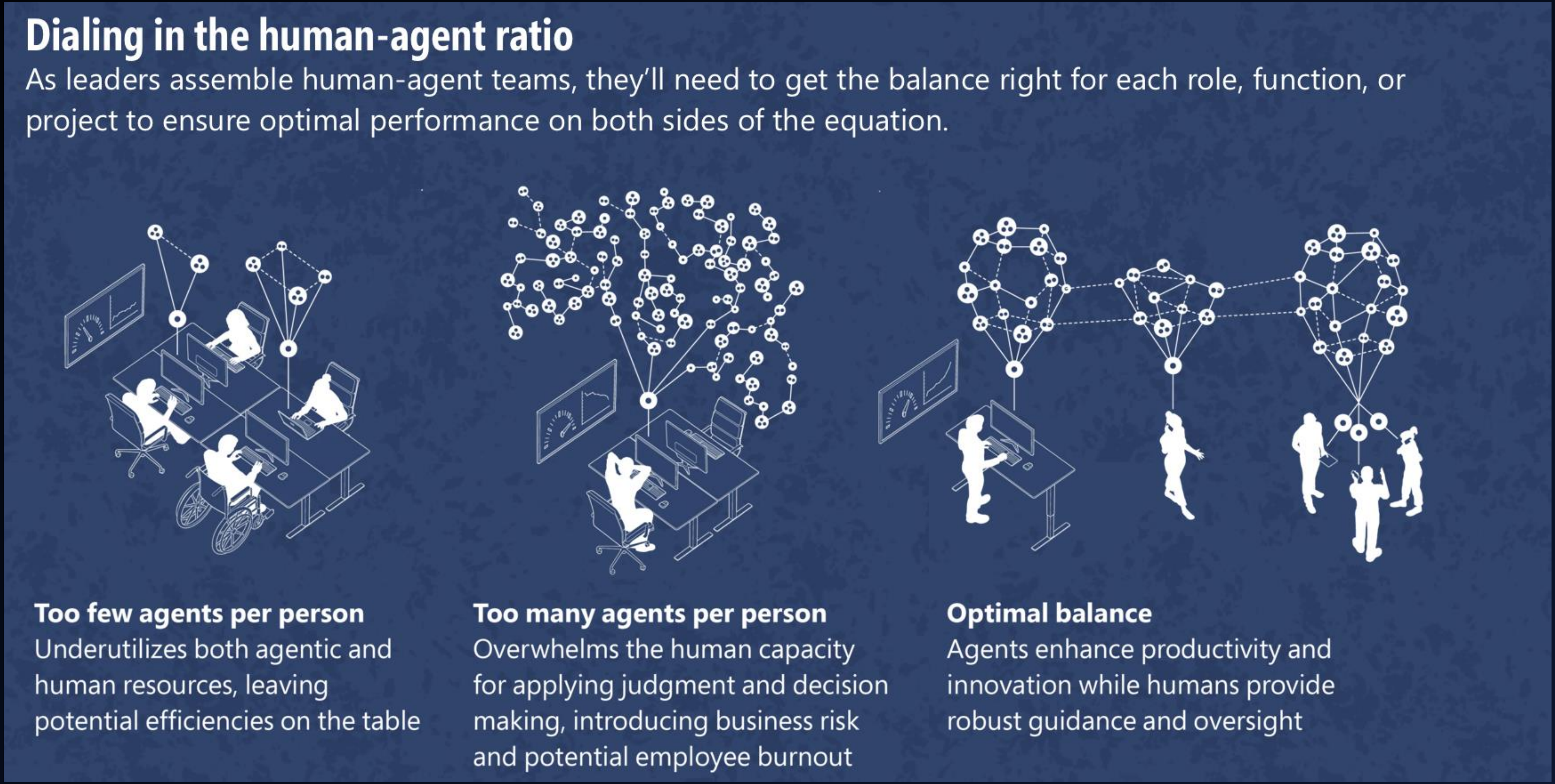
SOC Use case: Incident response

- Assist in identification of an ongoing attack
- Assist in assessing scale of an attack
- Assist in providing guidance around remediation

SOC Use case: Security reporting

- Assist in easily summarizing an event, incident, or threat
- Assist with the preparation of information into shareable and customizable reports

Directionally where are we heading?



Dialing in the human-agent ratio

As leaders assemble human-agent teams, they'll need to get the balance right for each role, function, or project to ensure optimal performance on both sides of the equation.

Too few agents per person
Underutilizes both agentic and human resources, leaving potential efficiencies on the table

Too many agents per person
Overwhelms the human capacity for applying judgment and decision making, introducing business risk and potential employee burnout

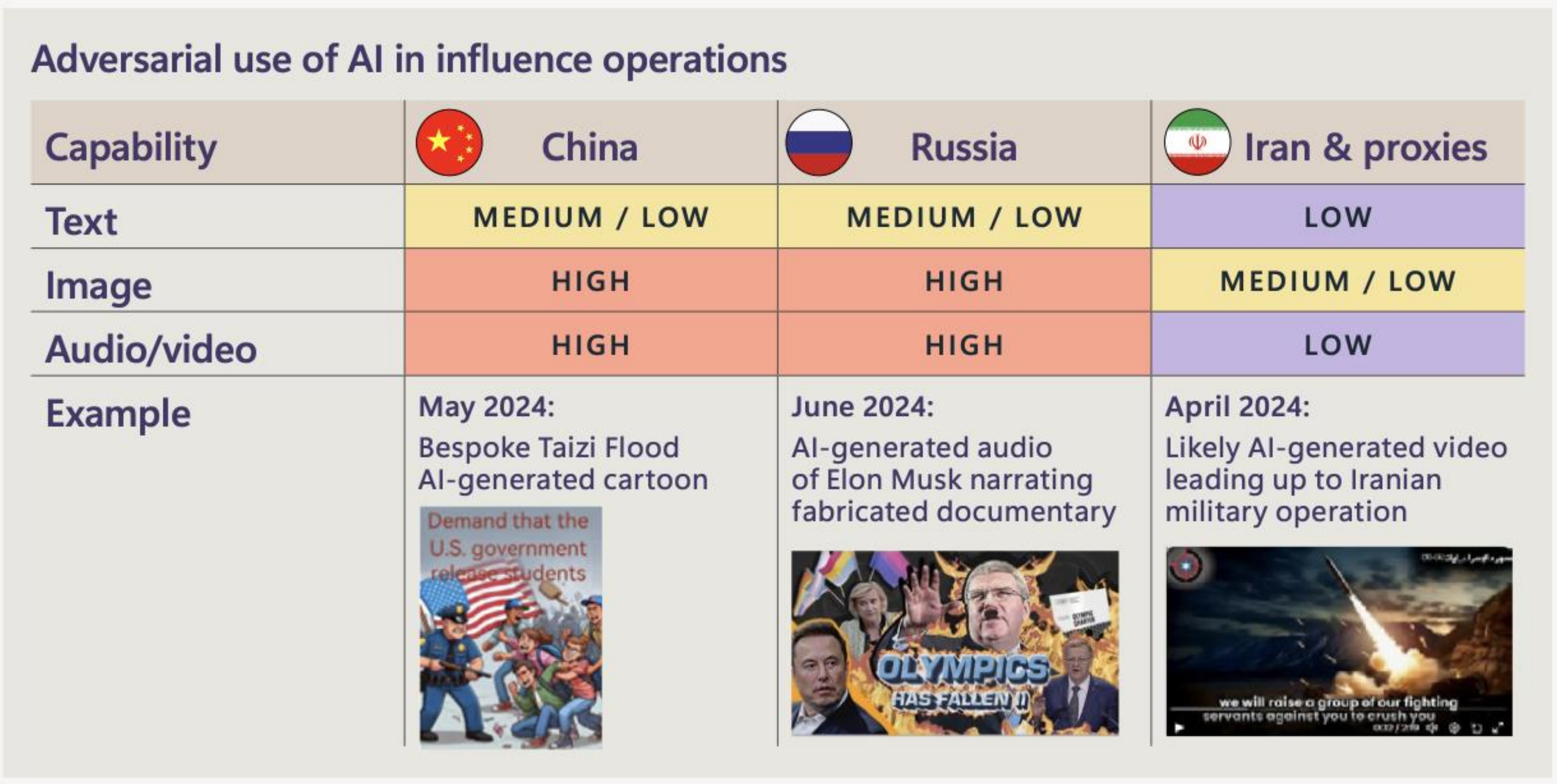
Optimal balance
Agents enhance productivity and innovation while humans provide robust guidance and oversight

Security of AI

AI can be used by bad actors with malicious intent such as criminals, terrorists, and hostile nation-states.

- Deep fakes
- Disinformation campaigns
- Misuse of military robots
- Autonomous weapon systems
- Social engineering
- Hacking and cyber attacks

Adversaria use of AI



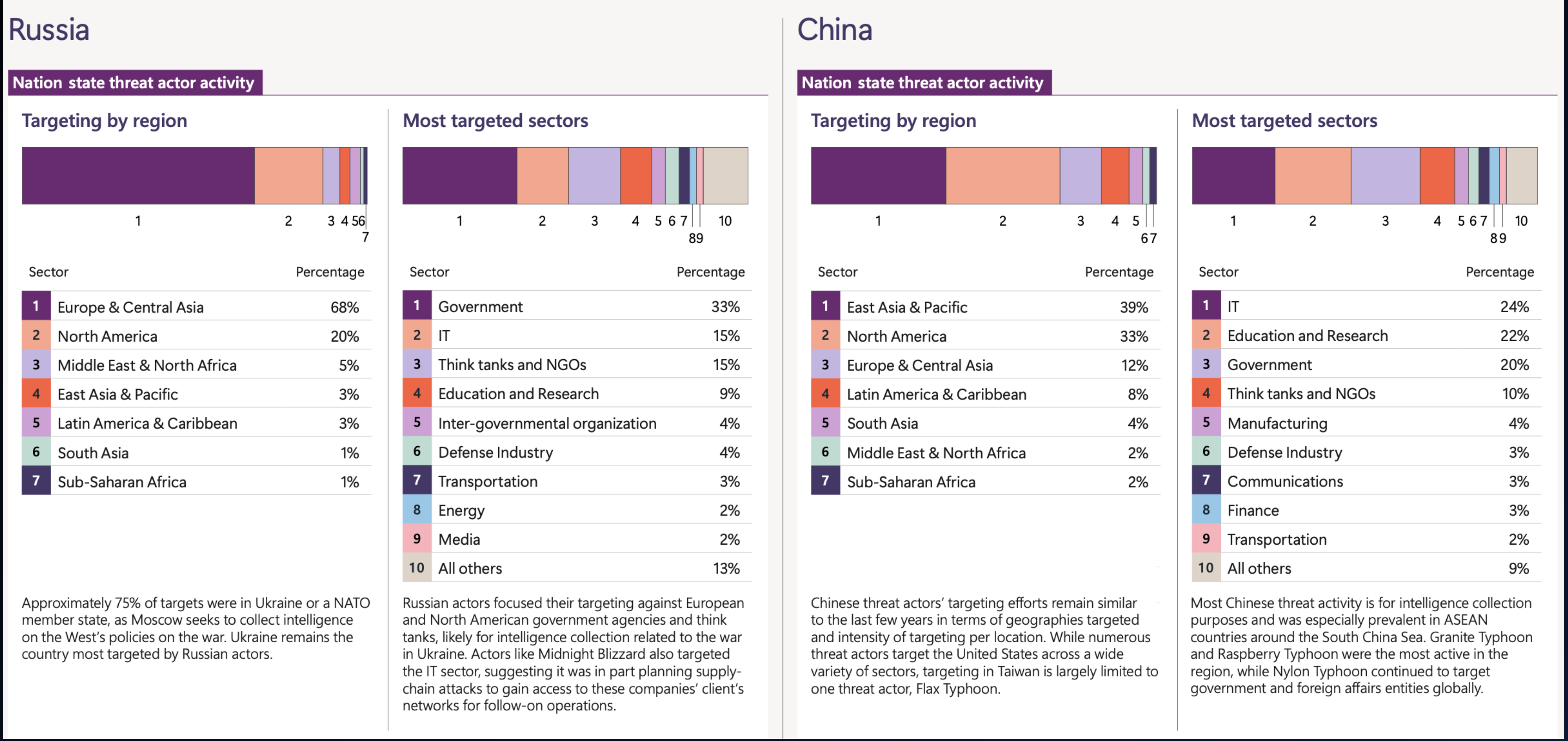
Adversarial use of AI in influence operations

Capability	China	Russia	Iran & proxies
Text	MEDIUM / LOW	MEDIUM / LOW	LOW
Image	HIGH	HIGH	MEDIUM / LOW
Audio/video	HIGH	HIGH	LOW
Example	May 2024: Bespoke Taizi Flood AI-generated cartoon	June 2024: AI-generated audio of Elon Musk narrating fabricated documentary	April 2024: Likely AI-generated video leading up to Iranian military operation

Demand that the U.S. government release students
OLYMPICS HAS FALLEN !!
we will raise a group of our fighting servants against you to crush you

Microsoft Digital Defense 2024

Nation State Actors & Targeted Sectors



Russia

Nation state threat actor activity

Targeting by region

	Sector	Percentage
1	Europe & Central Asia	68%
2	North America	20%
3	Middle East & North Africa	5%
4	East Asia & Pacific	3%
5	Latin America & Caribbean	3%
6	South Asia	1%
7	Sub-Saharan Africa	1%

Approximately 75% of targets were in Ukraine or a NATO member state, as Moscow seeks to collect intelligence on the West's policies on the war. Ukraine remains the country most targeted by Russian actors.

Most targeted sectors

	Sector	Percentage
1	Government	33%
2	IT	15%
3	Think tanks and NGOs	15%
4	Education and Research	9%
5	Inter-governmental organization	4%
6	Defense Industry	4%
7	Transportation	3%
8	Energy	2%
9	Media	2%
10	All others	13%

Russian actors focused their targeting against European and North American government agencies and think tanks, likely for intelligence collection related to the war in Ukraine. Actors like Midnight Blizzard also targeted the IT sector, suggesting it was in part planning supply-chain attacks to gain access to these companies' client's networks for follow-on operations.

China

Nation state threat actor activity

Targeting by region

	Sector	Percentage
1	East Asia & Pacific	39%
2	North America	33%
3	Europe & Central Asia	12%
4	Latin America & Caribbean	8%
5	South Asia	4%
6	Middle East & North Africa	2%
7	Sub-Saharan Africa	2%

Chinese threat actors' targeting efforts remain similar to the last few years in terms of geographies targeted and intensity of targeting per location. While numerous threat actors target the United States across a wide variety of sectors, targeting in Taiwan is largely limited to one threat actor, Flax Typhoon.

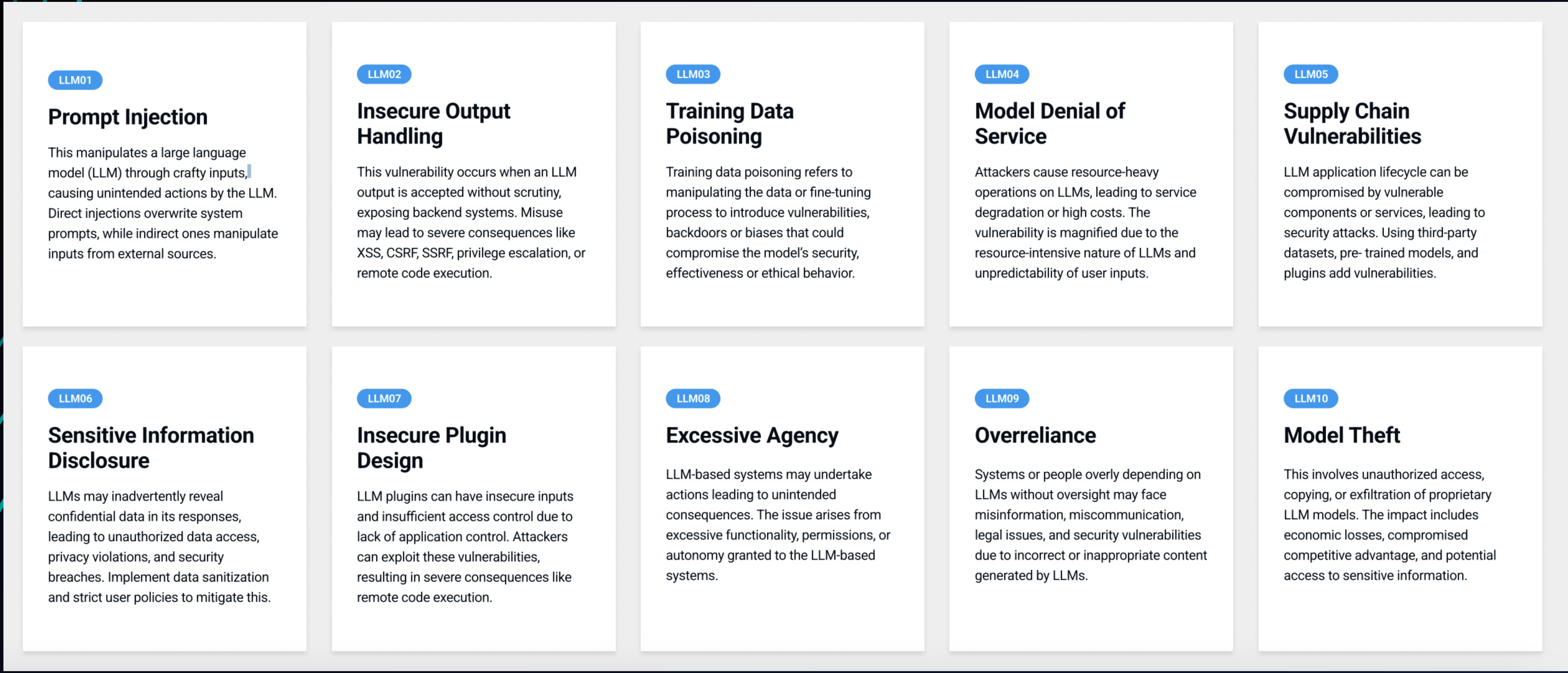
Most targeted sectors

	Sector	Percentage
1	IT	24%
2	Education and Research	22%
3	Government	20%
4	Think tanks and NGOs	10%
5	Manufacturing	4%
6	Defense Industry	3%
7	Communications	3%
8	Finance	3%
9	Transportation	2%
10	All others	9%

Most Chinese threat activity is for intelligence collection purposes and was especially prevalent in ASEAN countries around the South China Sea. Granite Typhoon and Raspberry Typhoon were the most active in the region, while Nylon Typhoon continued to target government and foreign affairs entities globally.

Microsoft Digital Defense 2024

OWASP: Top 10 for LLM



LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

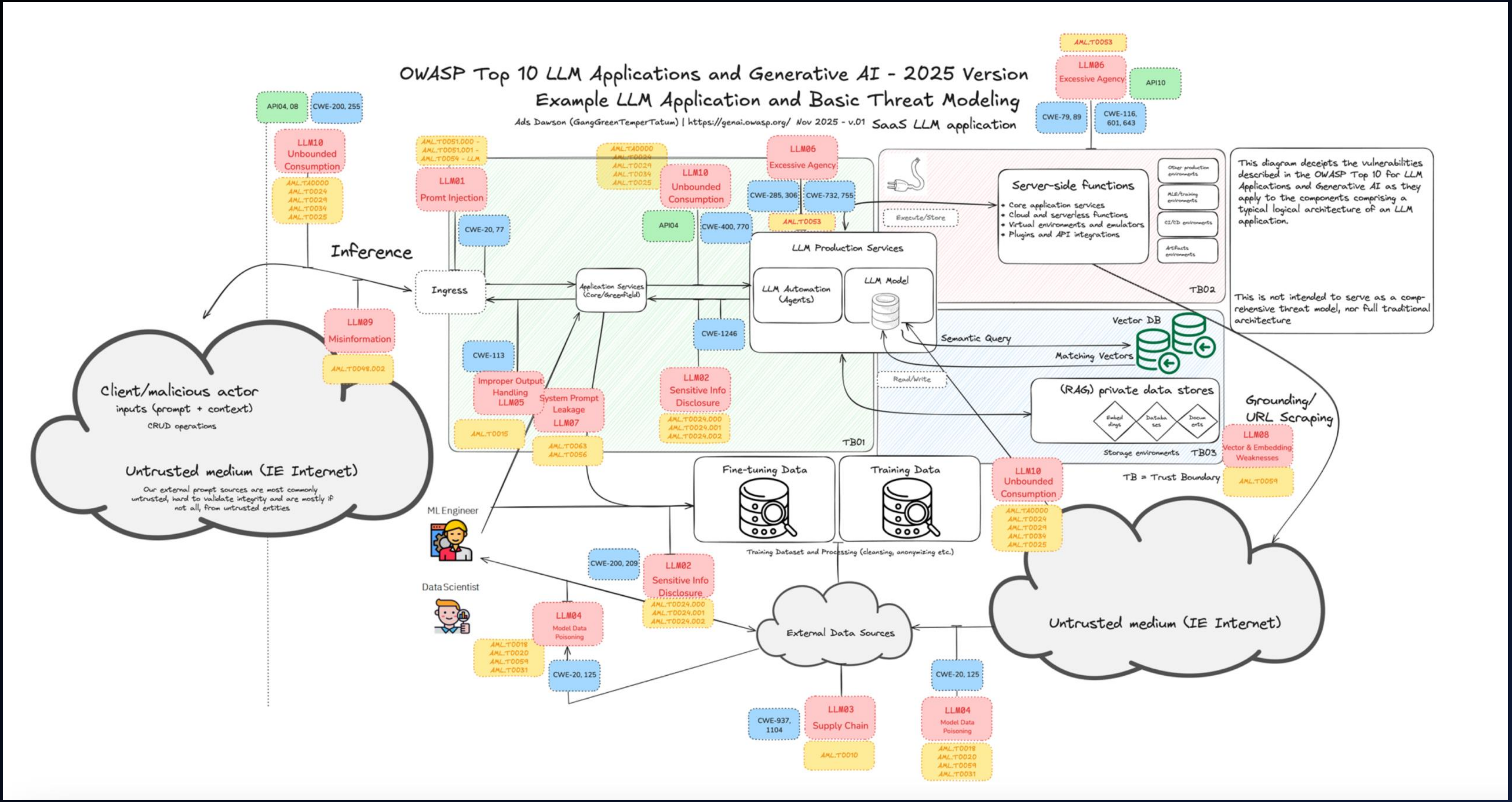
LLM10

Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

Ref: https://owasp.org/www-project-top-10-for-large-language-model-applications/

OWASP: Top 10 for LLM



OWASP Top 10 LLM Applications and Generative AI - 2025 Version

Example LLM Application and Basic Threat Modeling

Ads Dawson (GangGreenTemperTatum) | https://genai.owasp.org/ Nov 2025 - v.01 SaaS LLM application

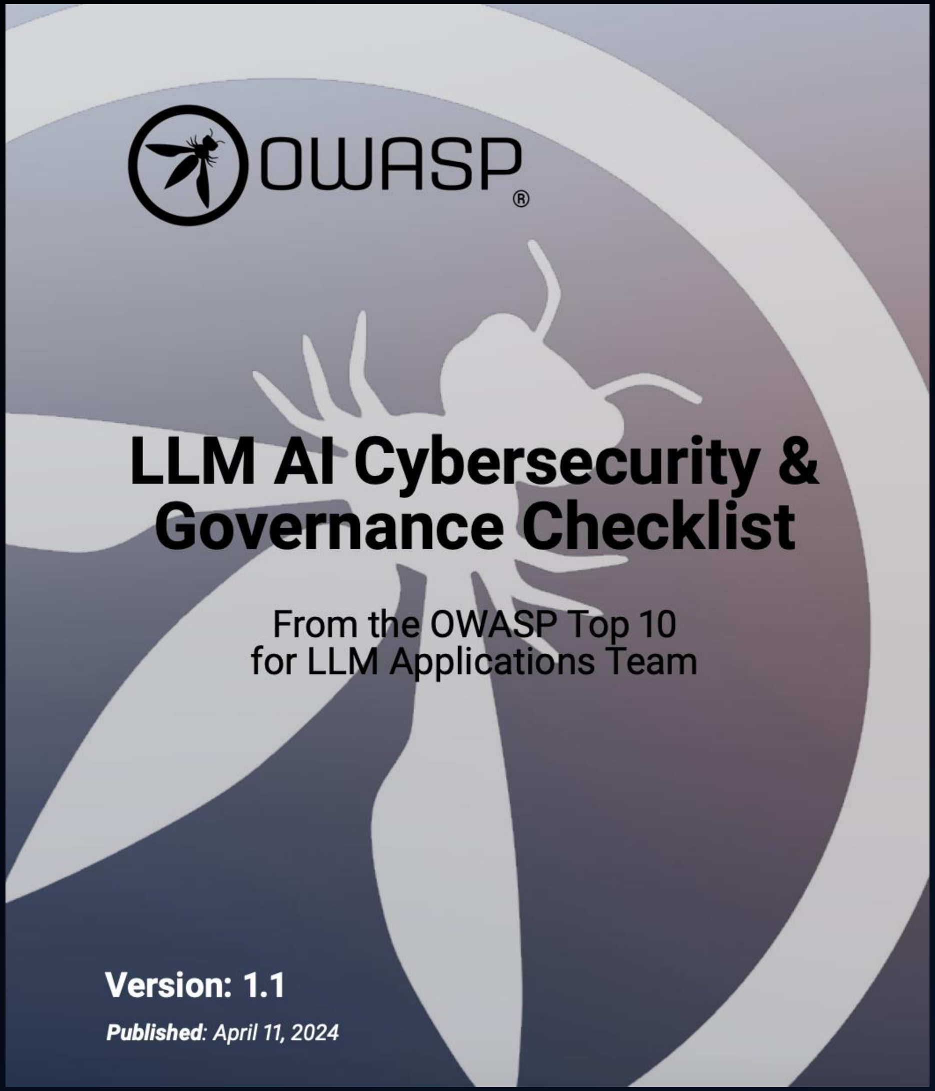
This diagram deceipts the vulnerabilities described in the OWASP Top 10 for LLM Applications and Generative AI as they apply to the components comprising a typical logical architecture of an LLM application.

This is not intended to serve as a comprehensive threat model, nor full traditional architecture

Our external prompt sources are most commonly untrusted, hard to validate integrity and are mostly if not all, from untrusted entities

TB = Trust Boundary

OWASP: LLM AI Cybersecurity & Governance Checklist



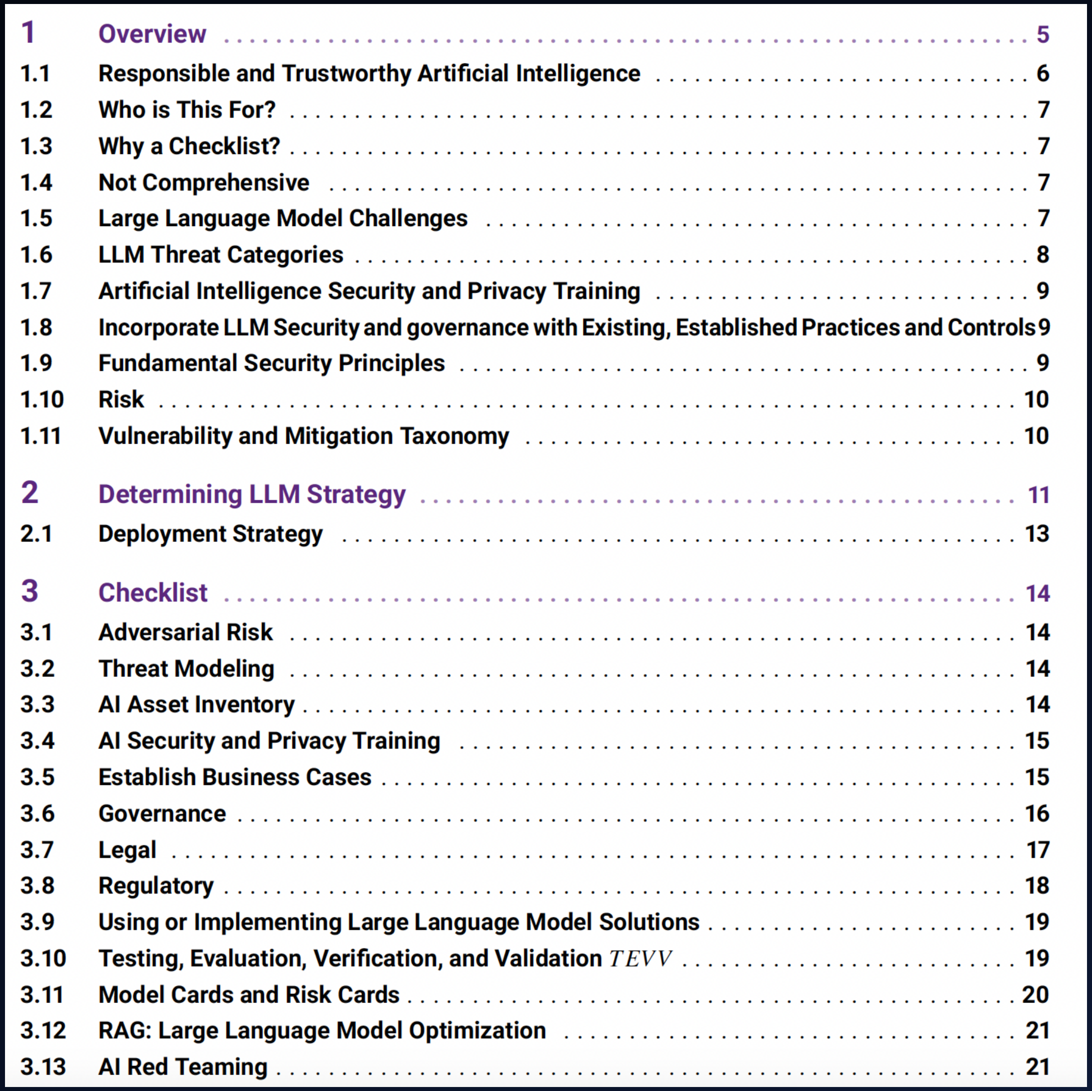
OWASP®

LLM AI Cybersecurity & Governance Checklist

From the OWASP Top 10 for LLM Applications Team

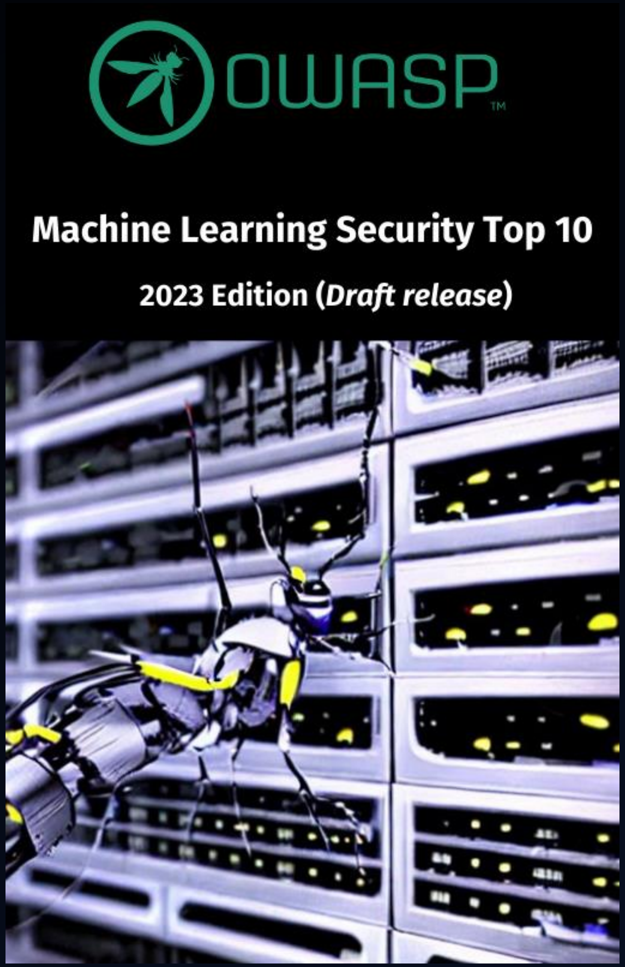
Version: 1.1

***Published**: April 11, 2024*



Ref: https://atlas.mitre.org/matrices/ATLAS

OWASP: Other AI Resources



OWASP

Machine Learning Security Top 10

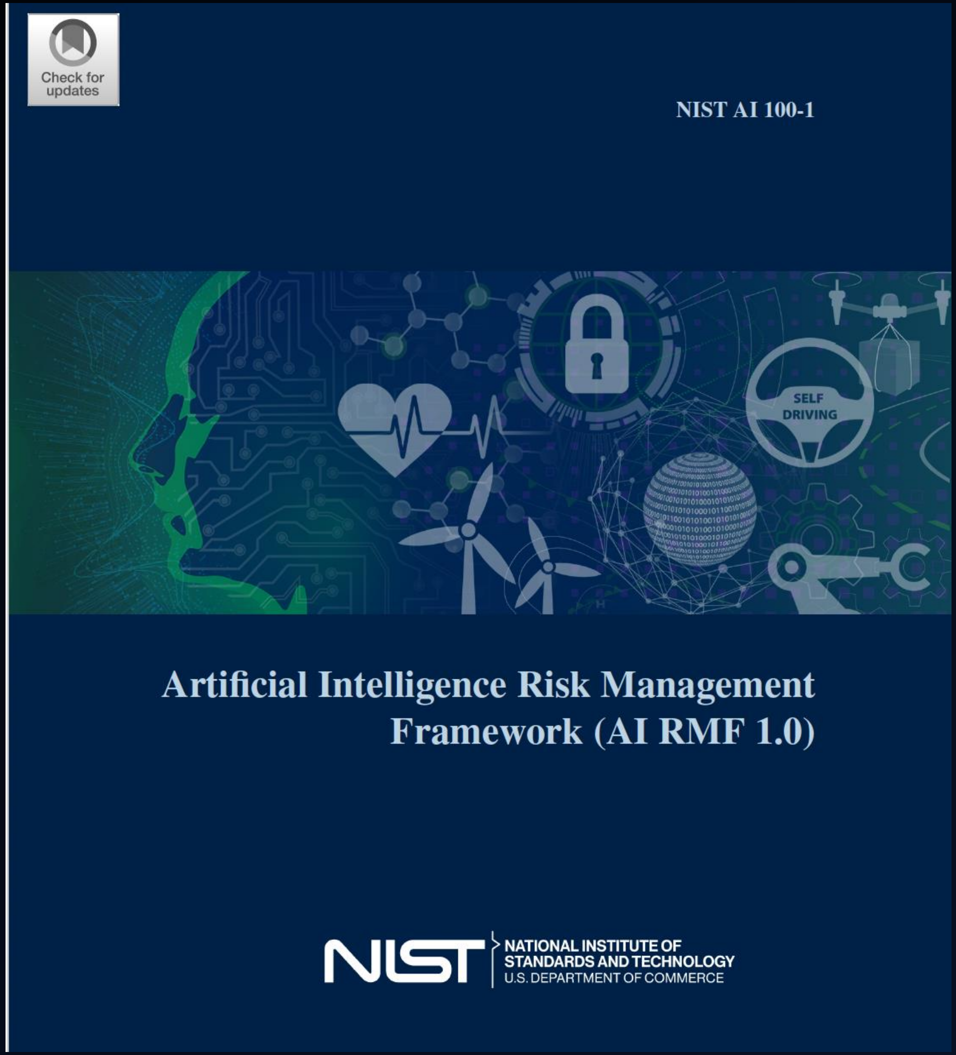
2023 Edition (*Draft release*)

https://mltop10.info/

https://owaspai.org/

https://owasp.org/www-project-ai-security-and-privacy-guide/

NIST: AI Risk Framework & Adversarial ML



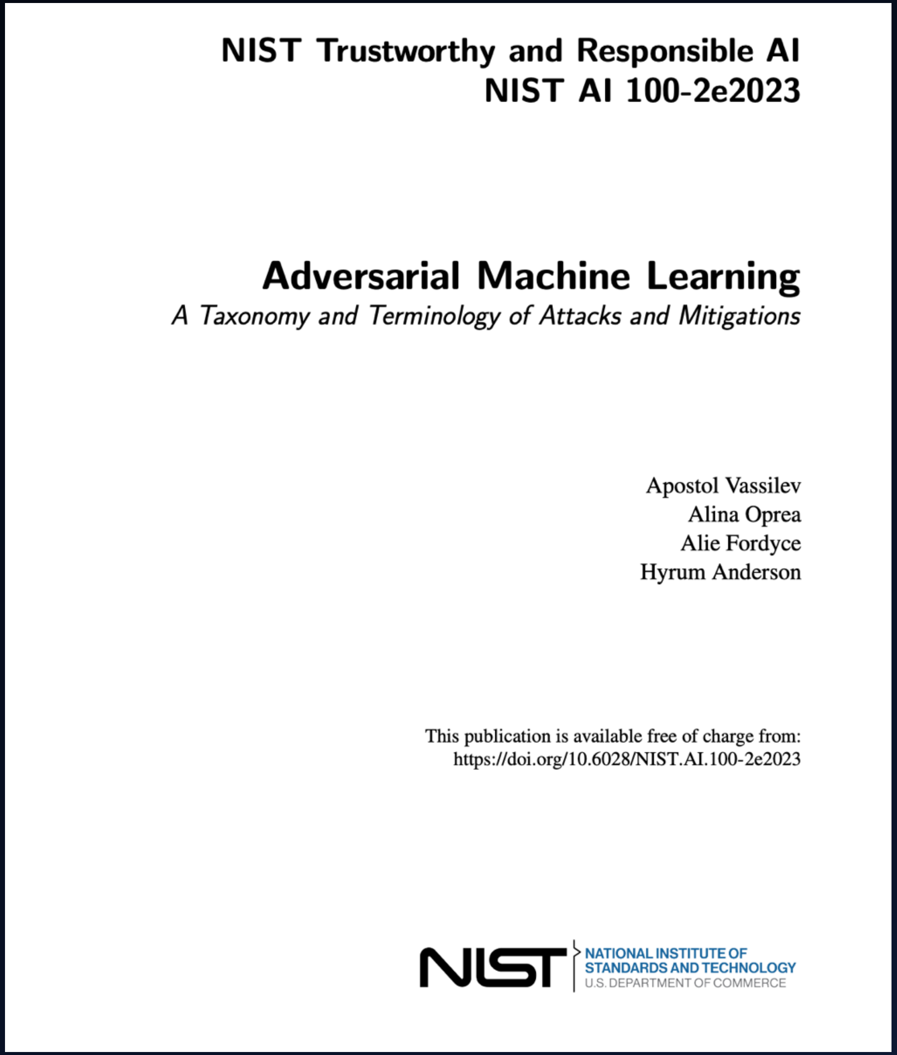
Check for updates

NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

https://www.nist.gov/itl/ai-risk-management-framework



NIST Trustworthy and Responsible AI
NIST AI 100-2e2023

Adversarial Machine Learning

A Taxonomy and Terminology of Attacks and Mitigations

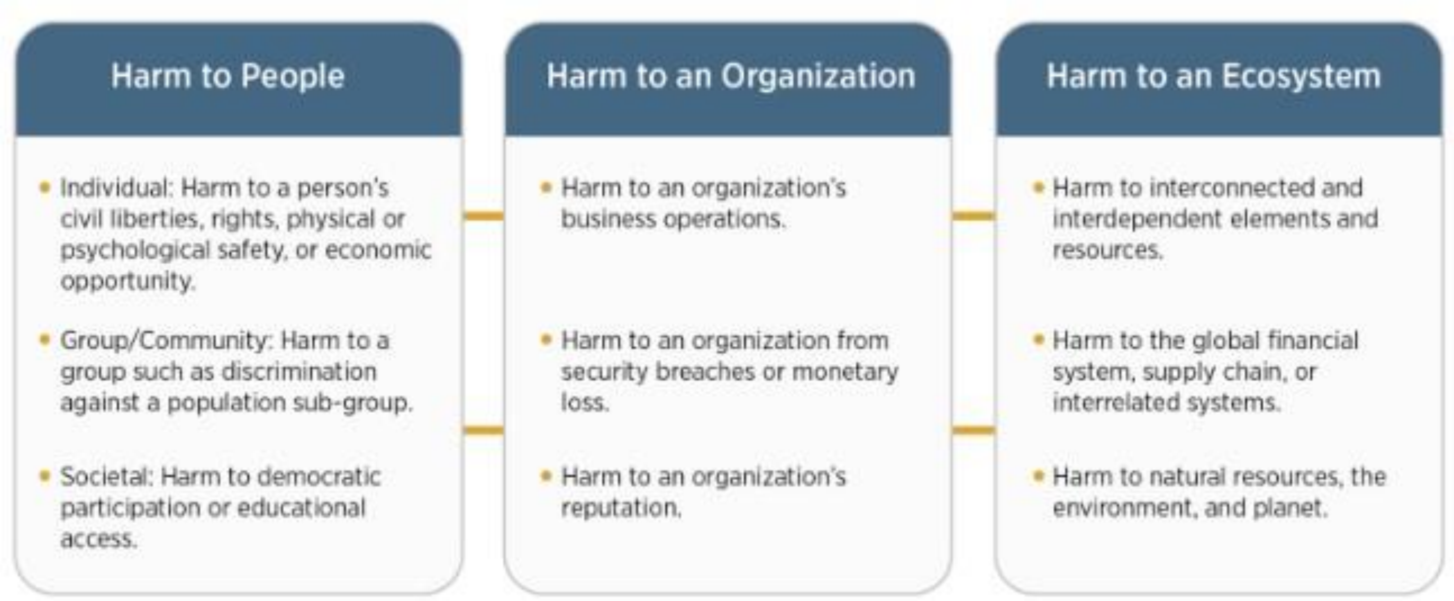
Apostol Vassilev
Alina Oprea
Alie Fordyce
Hyrum Anderson

This publication is available free of charge from:
https://doi.org/10.6028/NIST.AI.100-2e2023

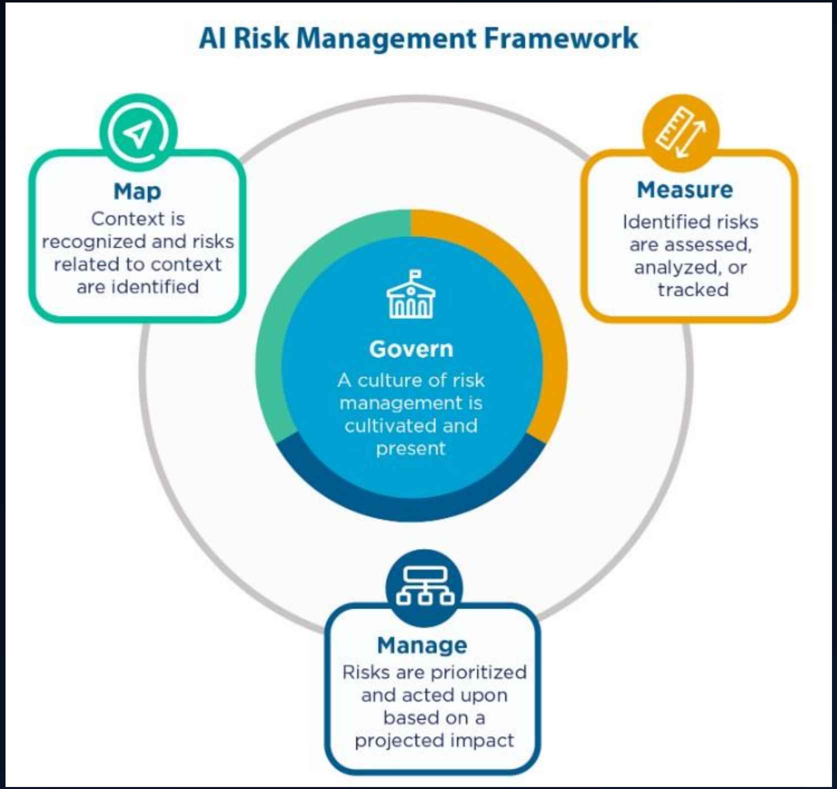
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

https://csrc.nist.gov/pubs/ai/100/2/e2023/final

NIST: AI Risk Framework



Harm to People	Harm to an Organization	Harm to an Ecosystem
Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity.	Harm to an organization's business operations.	Harm to interconnected and interdependent elements and resources.
Group/Community: Harm to a group such as discrimination against a population sub-group.	Harm to an organization from security breaches or monetary loss.	Harm to the global financial system, supply chain, or interrelated systems.
Societal: Harm to democratic participation or educational access.	Harm to an organization's reputation.	Harm to natural resources, the environment, and planet.



AI Risk Management Framework

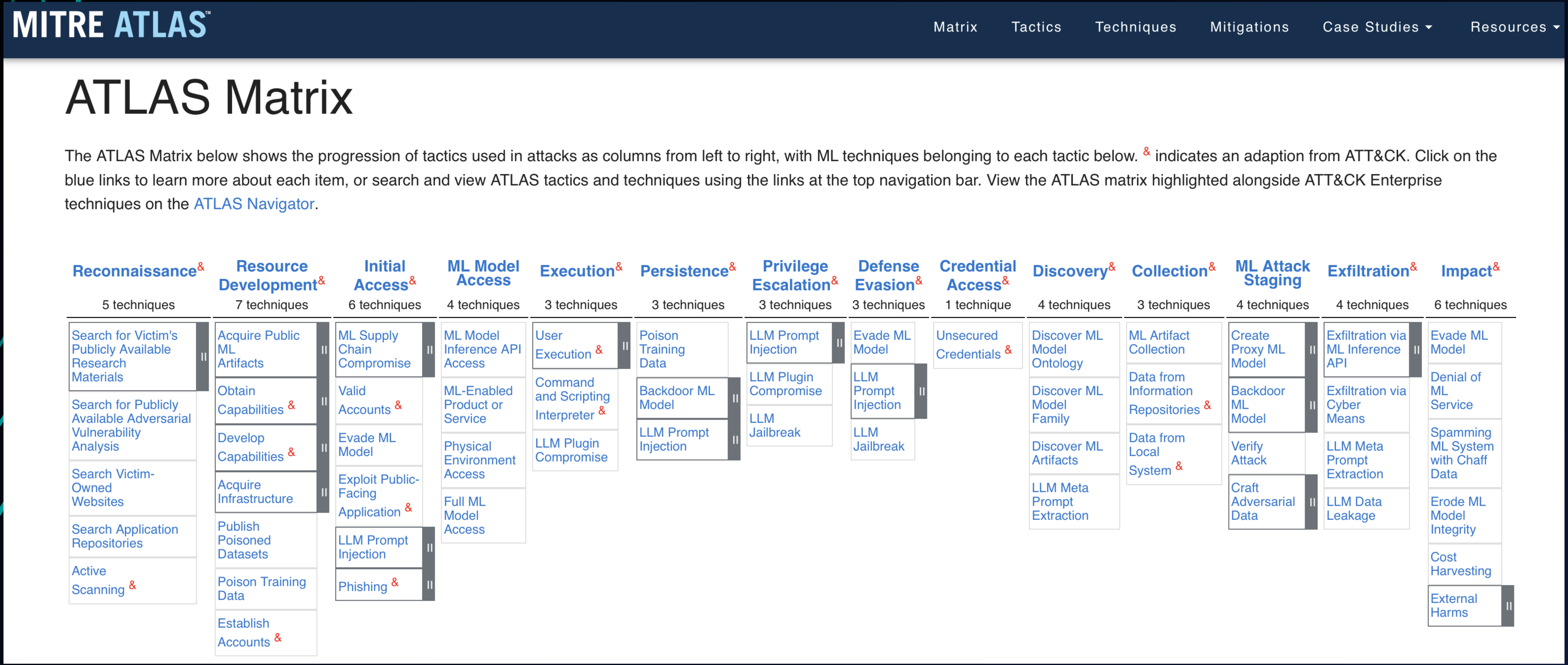
Map
Context is recognized and risks related to context are identified

Measure
Identified risks are assessed, analyzed, or tracked

Govern
A culture of risk management is cultivated and present

Manage
Risks are prioritized and acted upon based on a projected impact

MITRE: Adversarial Threat Landscape for AI Systems (ATLAS™)



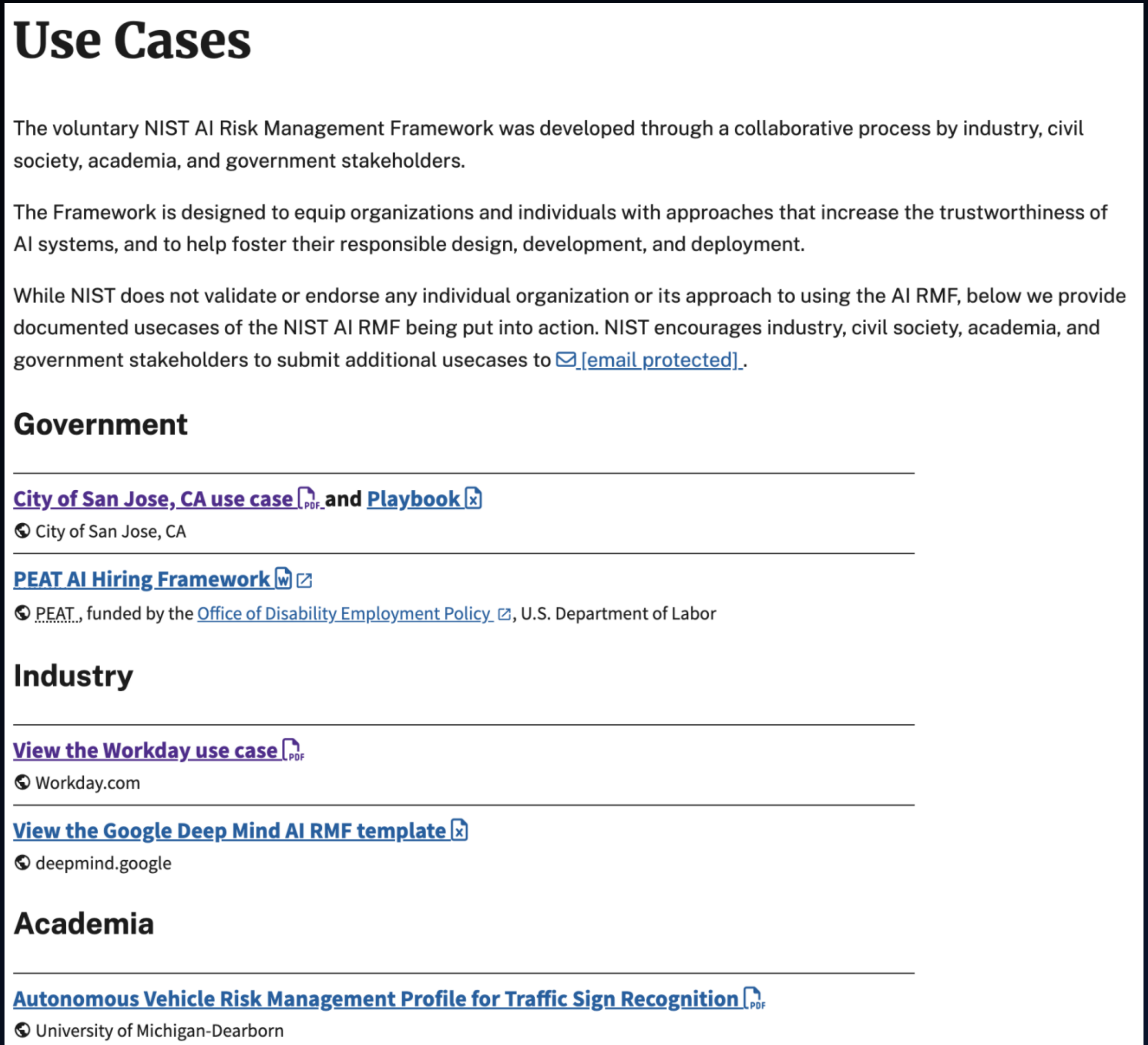
ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

Ref: https://atlas.mitre.org/matrices/ATLAS

NIST: AI Risk Framework Use Cases



Use Cases

The voluntary NIST AI Risk Management Framework was developed through a collaborative process by industry, civil society, academia, and government stakeholders.

The Framework is designed to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster their responsible design, development, and deployment.

While NIST does not validate or endorse any individual organization or its approach to using the AI RMF, below we provide documented usecases of the NIST AI RMF being put into action. NIST encourages industry, civil society, academia, and government stakeholders to submit additional usecases to [email protected].

Government

City of San Jose, CA use case and Playbook

City of San Jose, CA

PEAT AI Hiring Framework

PEAT, funded by the Office of Disability Employment Policy, U.S. Department of Labor

Industry

View the Workday use case

Workday.com

View the Google Deep Mind AI RMF template

deepmind.google

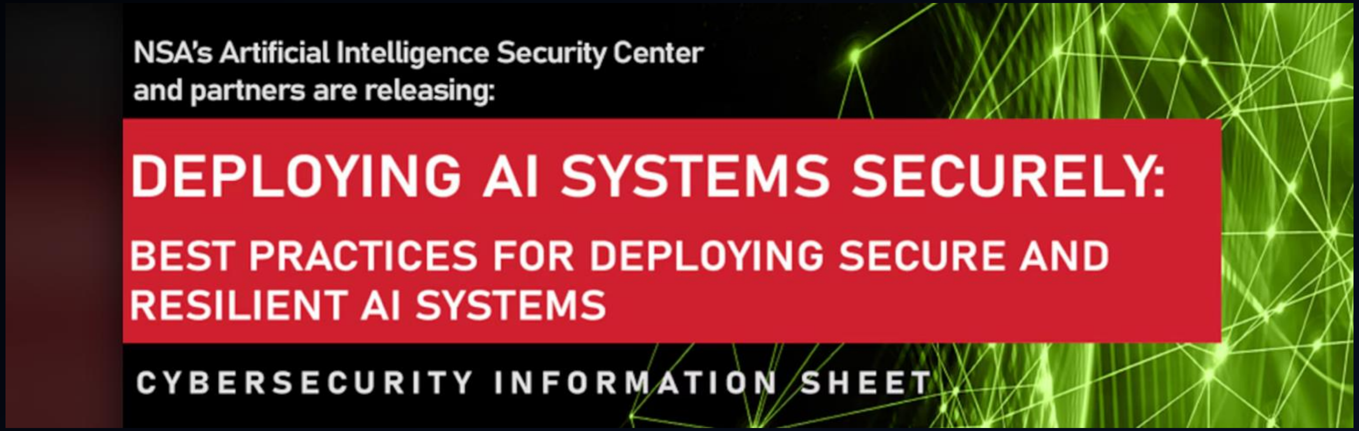
Academia

Autonomous Vehicle Risk Management Profile for Traffic Sign Recognition

University of Michigan-Dearborn

https://airc.nist.gov/airmf-resources/usecases/

NSA - Guidance for Strengthening AI System Security



NSA's Artificial Intelligence Security Center and partners are releasing:

DEPLOYING AI SYSTEMS SECURELY:

BEST PRACTICES FOR DEPLOYING SECURE AND RESILIENT AI SYSTEMS

CYBERSECURITY INFORMATION SHEET

Joint Cybersecurity Information

TLP:CLEAR

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

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

https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3741371/nsa-publishes-guidance-for-strengthening-ai-system-security/

Artificial Intelligence Engineer Certificate by AETIBA

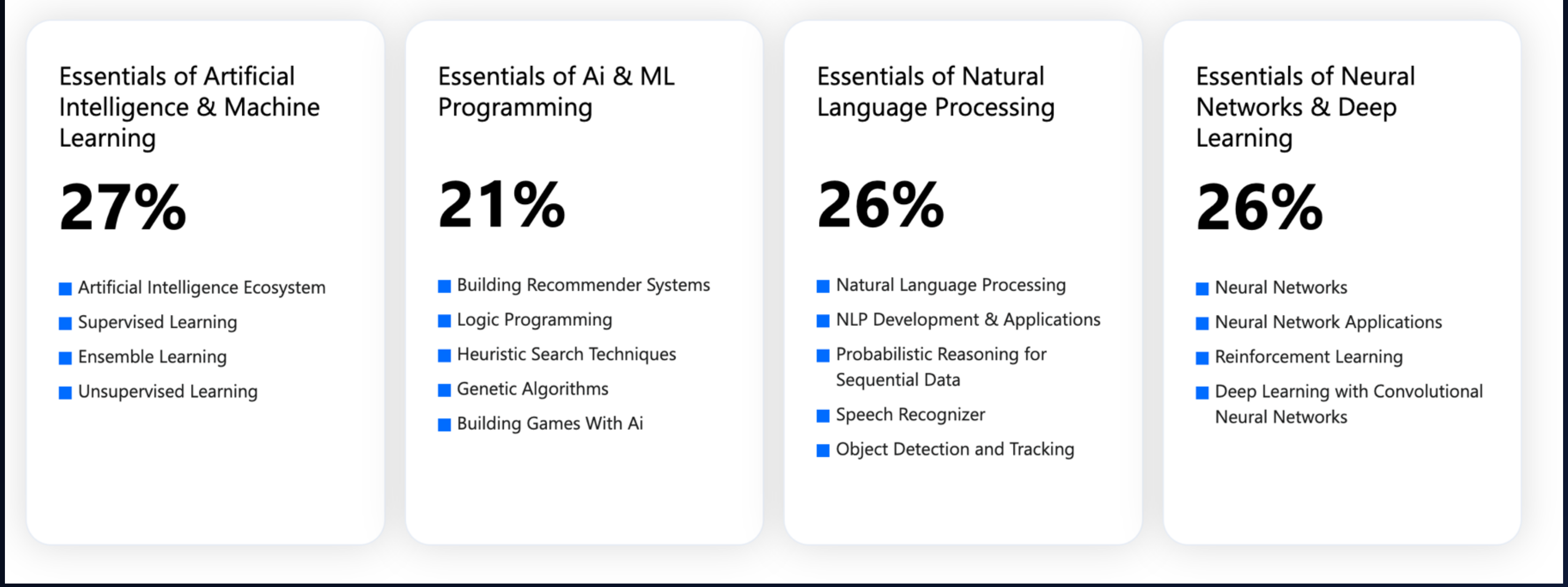
AI Engineers

AiE™

ARTIFICIAL INTELLIGENCE ENGINEER

https://www.artiba.org/certification/artificial-intelligence-certification

AiE by Artificial Intelligence Board of America (ARTIBA)



Essentials of Artificial Intelligence & Machine Learning

27%

- Artificial Intelligence Ecosystem
- Supervised Learning
- Ensemble Learning
- Unsupervised Learning

Essentials of Ai & ML Programming

21%

- Building Recommender Systems
- Logic Programming
- Heuristic Search Techniques
- Genetic Algorithms
- Building Games With Ai

Essentials of Natural Language Processing

26%

- Natural Language Processing
- NLP Development & Applications
- Probabilistic Reasoning for Sequential Data
- Speech Recognizer
- Object Detection and Tracking

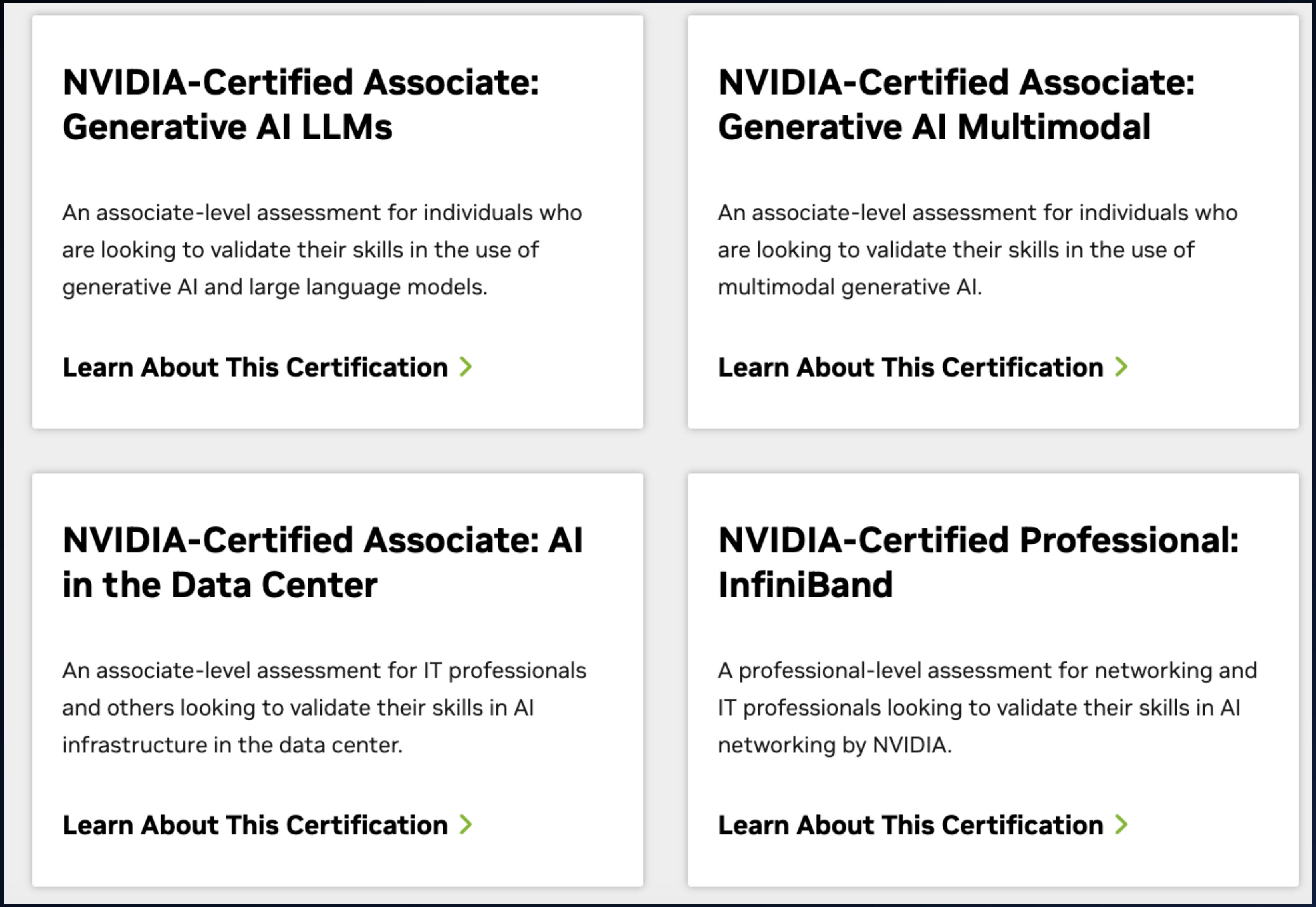
Essentials of Neural Networks & Deep Learning

26%

- Neural Networks
- Neural Network Applications
- Reinforcement Learning
- Deep Learning with Convolutional Neural Networks

NVIDIA AI Certifications

https://www.nvidia.com/en-us/learn/certification/



NVIDIA-Certified Associate: Generative AI LLMs

An associate-level assessment for individuals who are looking to validate their skills in the use of generative AI and large language models.

Learn About This Certification >

NVIDIA-Certified Associate: Generative AI Multimodal

An associate-level assessment for individuals who are looking to validate their skills in the use of multimodal generative AI.

Learn About This Certification >

NVIDIA-Certified Associate: AI in the Data Center

An associate-level assessment for IT professionals and others looking to validate their skills in AI infrastructure in the data center.

Learn About This Certification >

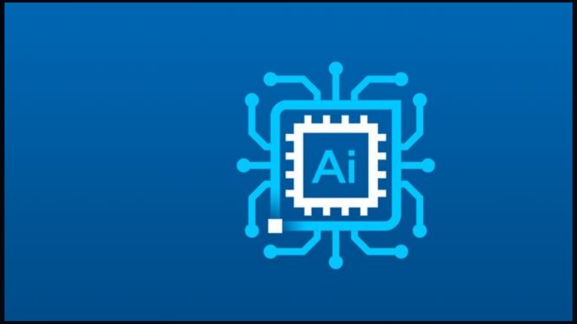
NVIDIA-Certified Professional: InfiniBand

A professional-level assessment for networking and IT professionals looking to validate their skills in AI networking by NVIDIA.

Learn About This Certification >

AI Certificate Courses

Intel Edge AI Certification



https://www.intel.com/content/www/us/en/developer/tools/devcloud/edge/learn/certification.html

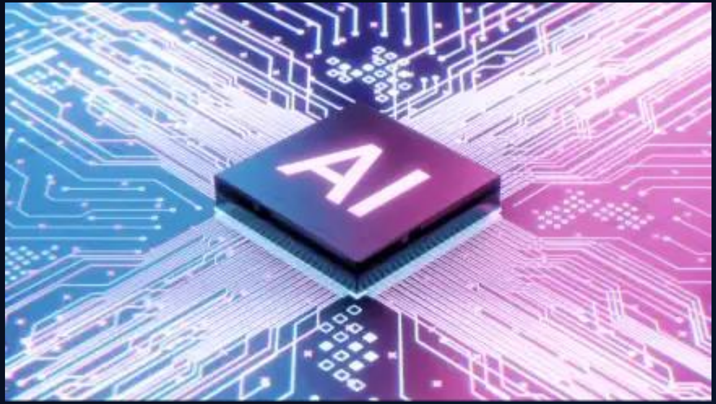
JETSON AI Courses and Certifications

NVIDIA.

DEEP LEARNING INSTITUTE

https://learn.microsoft.com/en-us/credentials/certifications/azure-ai-engineer/

edX



https://www.edx.org/learn/artificial-intelligence#browse-courses

Thank you!