Hacking AI
Risks and Rewards For Cybersecurity

AI Omaha Meetup
Feb 1, 2024

Ron Woerner

Warning!

- These are my thoughts based on my studies and experiences and NOT necessarily those of my employers or anyone else
- Ethical uses only
- Use at your own risk / Normal caveats apply
- There is homework!

PARENTS STRONGLY CAUTIONED
PG-13
SOME THEMATIC ELEMENTS, AND LANGUAGE
Some Material May Be Inappropriate for Children Under 13

whoami

Identity Paradox

How do you know
(I'm not a deep-fake)?

whoami

Ron Woerner[1]

- Hacker
- CyberSecurity Consultant / Trusted Advisor[2]
- Professor, Bellevue University
- 25+ years experience in IT / Security
- Blogger, writer, and podcaster

Websites & Social Media:
https://linktr.ee/cyberron

Slides available at https://github.com/hackerron/AI-Cybersecurity

LinkedIn: https://www.linkedin.com/in/ronwoerner/

[1] Who I'm claiming to be atm
[2] Can't say my employer

WHY

are we here?

AI Cybersecurity News

Google

ai cybersecurity news

All | News | Images | Videos | Books | More | Tools

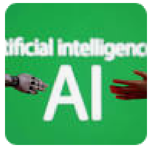
Recent | Sorted by relevance

Reuters

AI rise will lead to increase in cyberattacks, GCHQ warns

LONDON, Jan 24 (Reuters) - The rapid development of novel Artificial Intelligence (AI) tools will lead to an increase in cyberattacks and...

4 days ago



Security Magazine

Data Privacy Day 2024: Security leaders share AI concerns

With the ever-changing threat landscape, Data Privacy Day looks a little different each year as technology such as artificial intelligence...

1 day ago

TNW TechNewsWorld

AI in 2024 Brings Pivotal Shifts in Cybersecurity Trends

AI and quantum computing are reshaping the cybersecurity landscape. Expect a mix of advanced threats and cutting-edge defenses in 2024.

2 days ago

Austin American-Statesman

University of Texas-San Antonio wants AI, data science college

UTSA announced an initiative to establish a new college focused on AI, cybersecurity and computer and data science.

1 day ago

TheNextWeb

States could already produce AI malware that evades detection

Discussions of artificial intelligence (AI) often swirl with mysticism regarding how an AI system functions. The reality is far more simple:

AI is a type of software system.

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

AMERICA'S CYBER DEFENSE AGENCY

https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception

AI Risks & Threats

Fig. 1. Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems.

NIST AI Risk Management Framework 1.0, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, p.5

National Cyber Security Centre
a part of GCHQ

The near-term impact of AI on the cyber threat

An NCSC assessment focusing on how AI will impact the efficacy of cyber operations and the implications for the cyber threat over the next two years.

24 January 2024

Key judgements

- Artificial intelligence (AI) will **almost certainly increase the volume and heighten the impact of cyber attacks** over the next two years. However, the **impact on the cyber threat will be uneven** (see table 1).
- The threat to 2025 comes from **evolution and enhancement of existing tactics, techniques and procedures** (TTPs).
- **All types of cyber threat actor** – state and non-state, skilled and less skilled – are already using AI, **to varying degrees.**
- AI provides **capability uplift in reconnaissance and social engineering**, almost certainly making both more effective, efficient, and harder to detect.
- More **sophisticated uses of AI in cyber operations** are highly likely to be restricted to threat actors with access to **quality training data, significant expertise (in both AI and cyber), and resources.** More advanced uses are unlikely to be realised before 2025.
- AI will almost certainly make cyber attacks against the UK more impactful because **threat actors will be able to analyse exfiltrated data faster and more effectively, and use it to train AI models.**
- AI lowers the barrier for novice cyber criminals, hackers-for-hire and hacktivists to carry out effective access and information gathering operations. This **enhanced access will likely contribute to the global ransomware threat** over the next two years.
- Moving towards 2025 and beyond, commoditisation of AI-enabled capability in criminal and commercial markets will almost certainly make **improved capability available to cyber crime and state actors.**

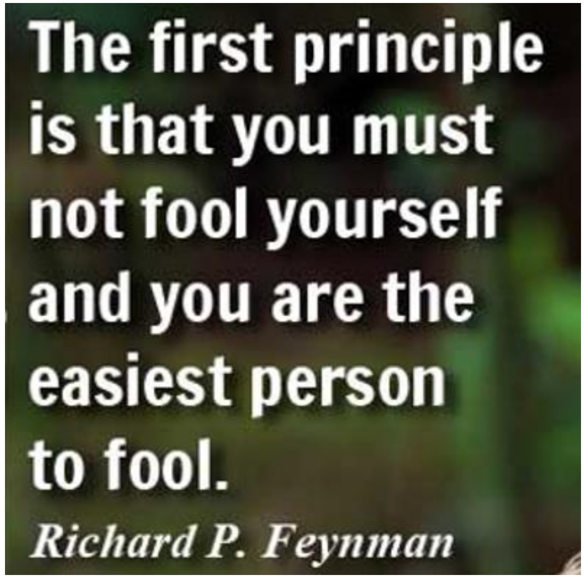
https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

AI Risks & Threats

Manipulating victims:
- Pissed off,
- Perturbed, or
- Panicked

Prompt: I'm building a presentation on cybersecurity and AI for a technical audience. Provide 5 ways ChatGPT and AI can be used maliciously.

- ➢ Automated social engineering
- ➢ Supercharged phishing
- ➢ Deepfakes and disinformation
- ➢ Enhanced cyber reconnaissance
- ➢ Malware automation & mutation



Malicious AI Examples

Automated Social Engineering:

- ***Problem***: AI-powered chatbots can engage in seemingly natural conversations, impersonating customer service representatives or trusted individuals to extract sensitive information.
- ***Impact***: Increased risk of identity theft, financial fraud, and data breaches by tricking users into revealing personal details.
- ***Example***: Imagine an AI chatbot posing as a bank representative calling you to "verify" your account information, eventually luring you into disclosing your PIN or credentials.

Supercharged Phishing:

- ***Problem***: ChatGPT excels at mimicking human writing styles and crafting personalized narratives. This makes AI-generated phishing emails more convincing and bypasses traditional spam filters.
- ***Impact***: Increased risk of sensitive data breaches, financial losses, and reputational damage for organizations.
- ***Example***: Imagine an email, Teams, Slack, etc. message seemingly from a trusted colleague praising your recent work and subtly prompting you to click a malicious link to access a “bonus document.”

Deepfakes and Disinformation:

- ***Problem***: AI can be used to manipulate audio and video to create realistic deepfakes that spread misinformation, damage reputations, and sow discord.
- ***Impact***: Erode trust in institutions, incite social unrest, and manipulate public opinion for nefarious purposes.
- ***Example***: A fabricated video portraying a political leader making inflammatory statements could go viral and disrupt democratic processes.

Malicious AI Examples

Enhanced Cyber Reconnaissance:

- *Problem*: AI can analyze vast amounts of data to identify vulnerabilities in networks, systems, and software, aiding attackers in targeting their efforts..
- *Impact*: Increased risk of successful cyberattacks as attackers gain valuable insights into potential entry points and exploit weaknesses.
- *Example*: Imagine an AI scouring open-source forums to find disgruntled employees mentioning security protocols, providing valuable intel for targeted attacks.

https://osintframework.com/

Malware Automation and Mutation:

- *Problem*: AI can be used to automate the creation and modification of malware, making it more sophisticated, evasive, and difficult to detect by traditional security measures.
- *Impact*: Increased risk of widespread malware outbreaks, data loss, and disruption of critical infrastructure.
- *Example*: Imagine an AI generating constantly evolving variants of a ransomware virus, making it nearly impossible to identify and neutralize before causing widespread damage.

WormGPT, https://flowgpt.com/p/wormgpt-6

Threat Modeling

1. What are we working on?
2. What can go wrong?
3. What are we going to do about it?
4. Did we do a good job?

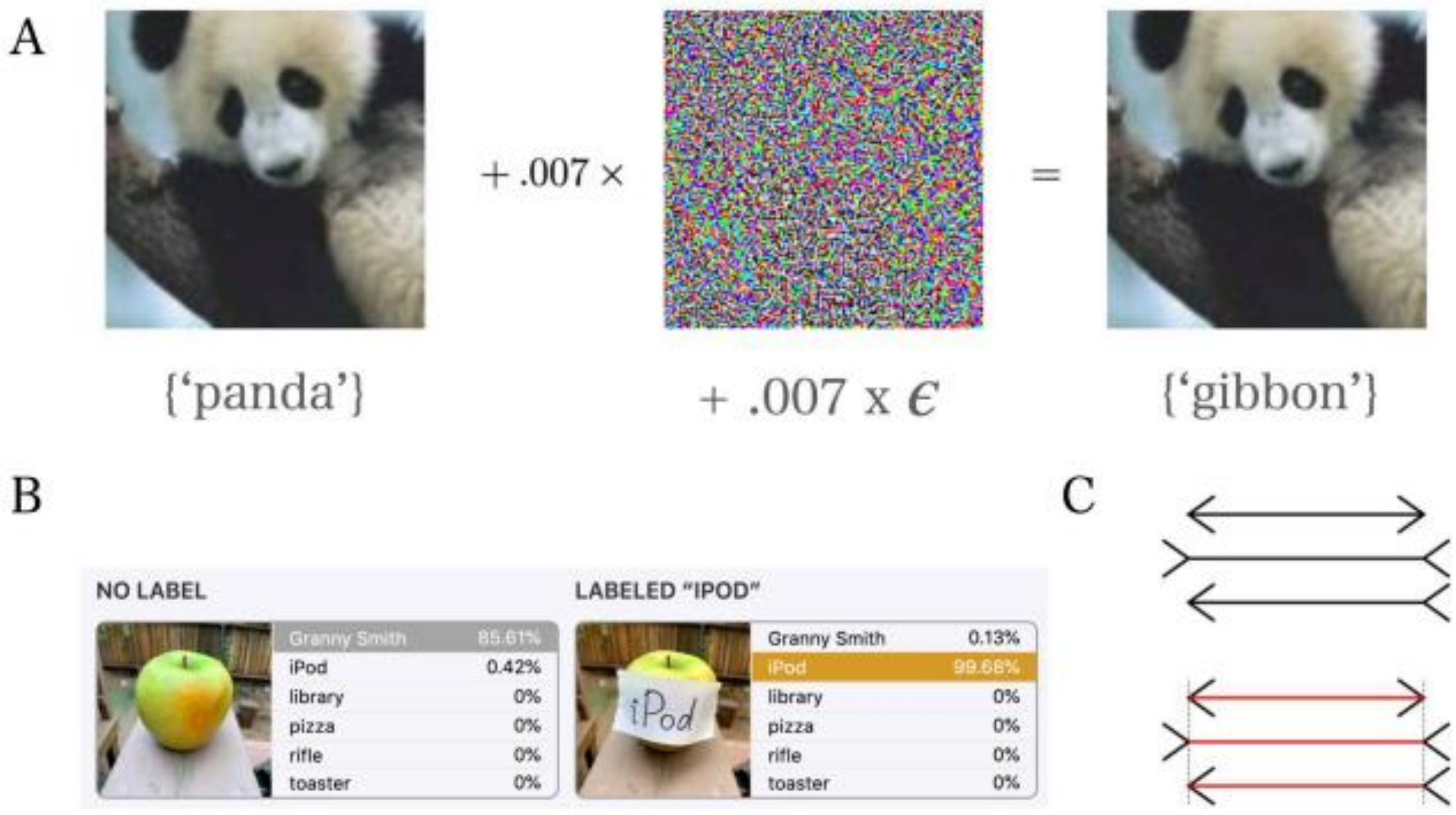
1. Shostack, A. (2014). *Threat modeling: Designing for security*. http://ci.nii.ac.jp/ncid/BB16065709
2. https://shostack.org/resources/threat-modeling
3. https://shostack.org/blog/category/threat-modeling
4. https://www.threatmodelingmanifesto.org/

Adversarial Machine Learning (AML)

- *Adversarial Machine Learning* (AML), is used to describe the **exploitation of fundamental vulnerabilities in ML components**, including hardware, software, workflows and supply chains.
- AML enables attackers to **cause unintended behaviours in ML systems** which can include:
  > affecting the model's classification or regression performance
  > allowing users to perform unauthorised actions
  > extracting sensitive model information
- Examples: **prompt injection** attacks in the large language model (LLM) domain, or **deliberately corrupting the training data** or user feedback (known as 'data poisoning').

Guidelines for Secure AI Development, p. 6,
https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf,
NCSC (UK) & CISA (US)

Adversarial Machine Learning



Source: Goodfellow et al., 2015 and Goh et al., 2021

ML Failure Modes – Intentional Failure

	Perturbation and Adversarial Universal Perturbation Attacks
	Poisoning Attacks
	Reprogramming Neural Nets
	3D Adversarial Objects
	Supply Chain Attacks
	Model Inversion
	Membership Inference and Model Stealing
	Backdoors and Existing Exploits

Kumar, et.al. (2022, November 2). *Failure modes in machine learning*. Microsoft Learn. https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

ML Failure Modes – Unintentional Failure

	Reward Hacking
	Side Effects
	Distributional Shifts
	Incomplete Testing
	Over/Under-Fitting
	Data Bias

Kumar, et.al. (2022, November 2). *Failure modes in machine learning*. Microsoft Learn. https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

MITRE | ATLAS™

ATLAS™

The ATLAS Matrix below shows the general progression of attack tactics as column headers from left to right, with attack techniques organized below each tactic. & indicates a tactic or technique directly adapted from from ATT&CK. Click on the blue links to learn more about each item, or search and view more details about ATLAS tactics and techniques using the links in the top navigation bar.

Reconnaissance & (5 techniques)	Resource Development & (7 techniques)	Initial Access & (6 techniques)	ML Model Access (4 techniques)	Execution & (3 techniques)	Persistence & (3 techniques)	Privilege Escalation & (3 techniques)	Defense Evasion & (3 techniques)	Credential Access & (1 technique)	Discovery & (4 techniques)	Collection & (3 techniques)	ML Attack Staging (4 techniques)	Exfiltration & (4 techniques)	Impact & (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 &												

https://atlas.mitre.org/

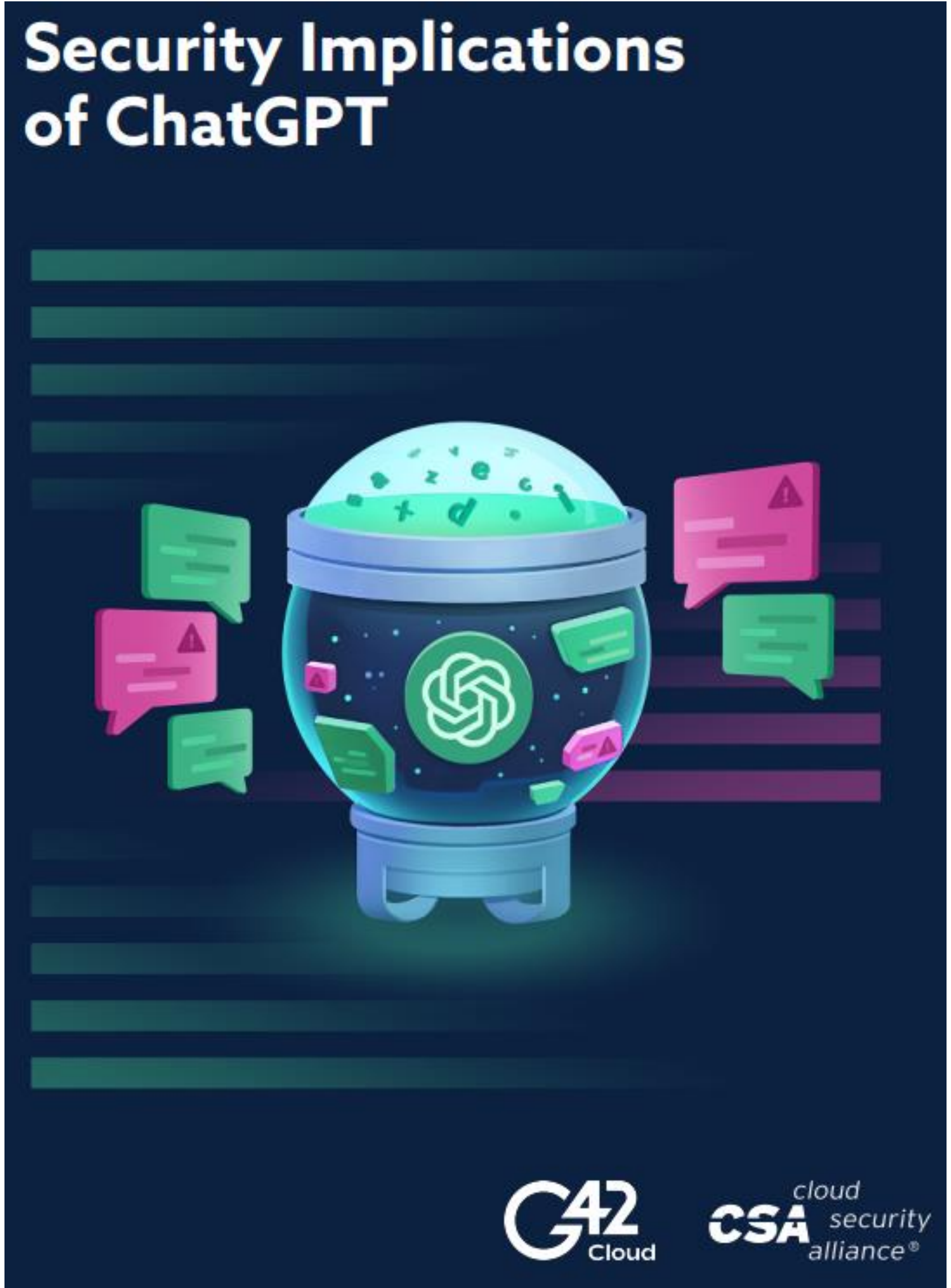
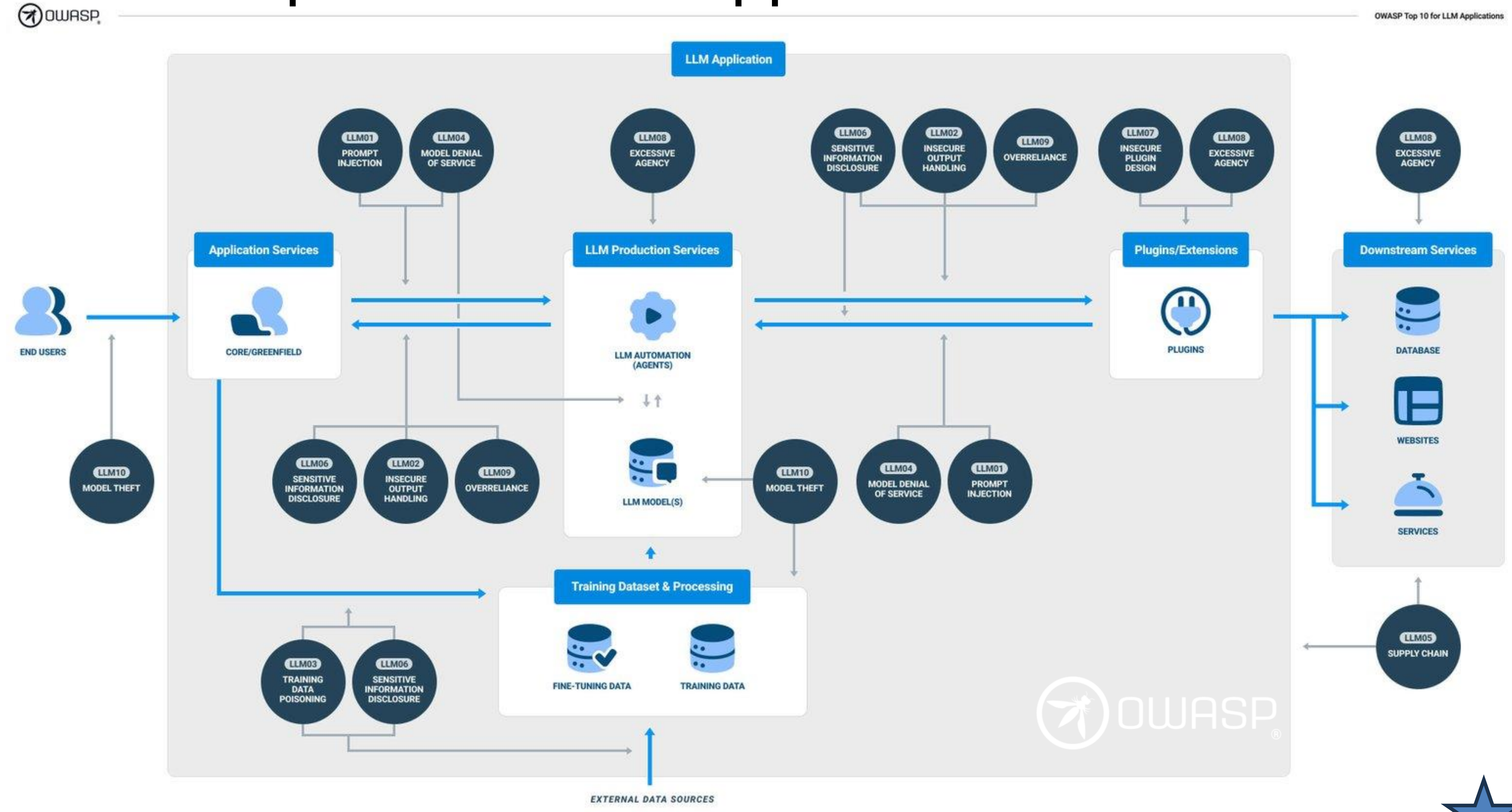


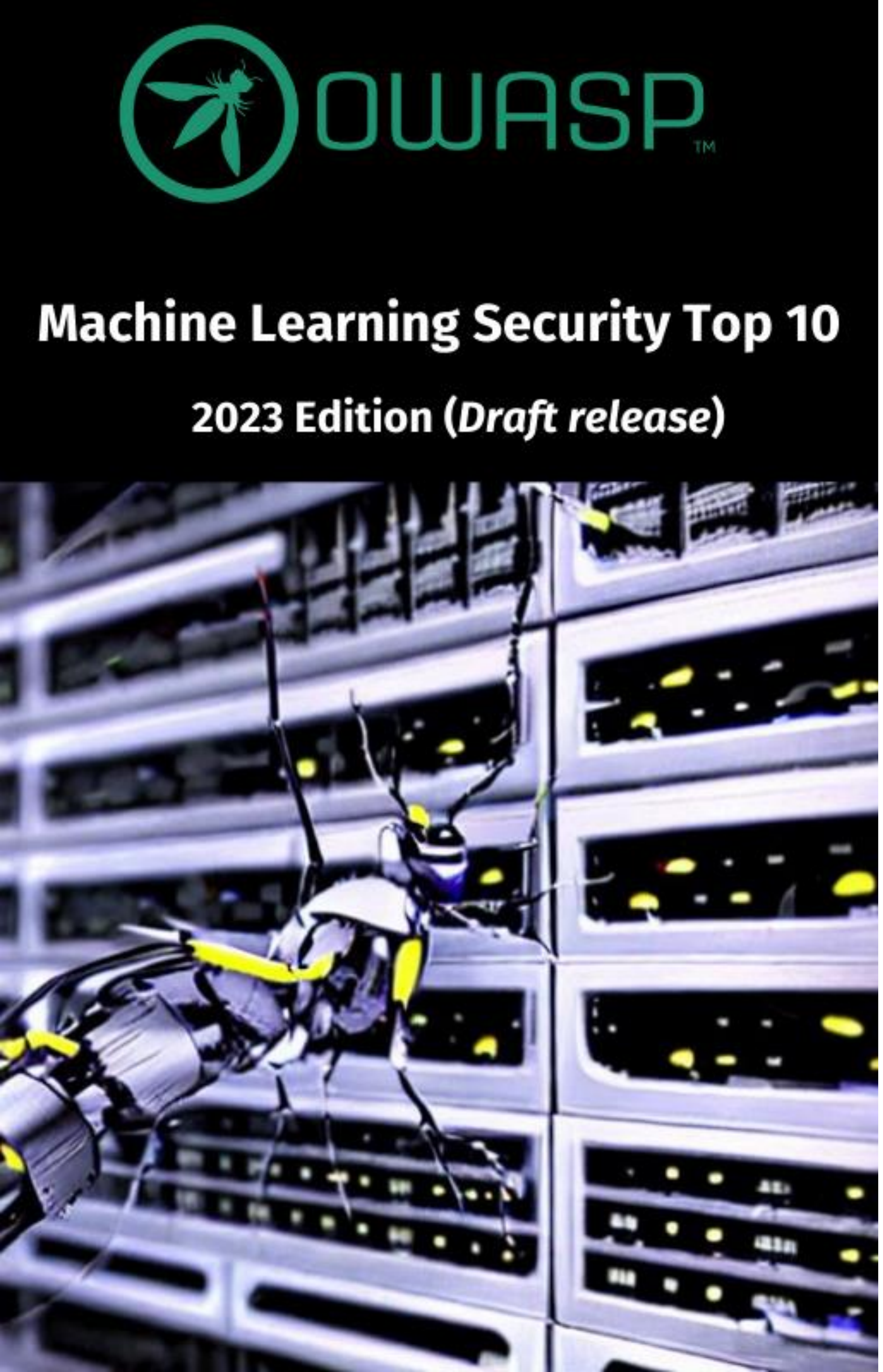
Table of Contents

https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/

OWASP Top 10 for LLM Applications



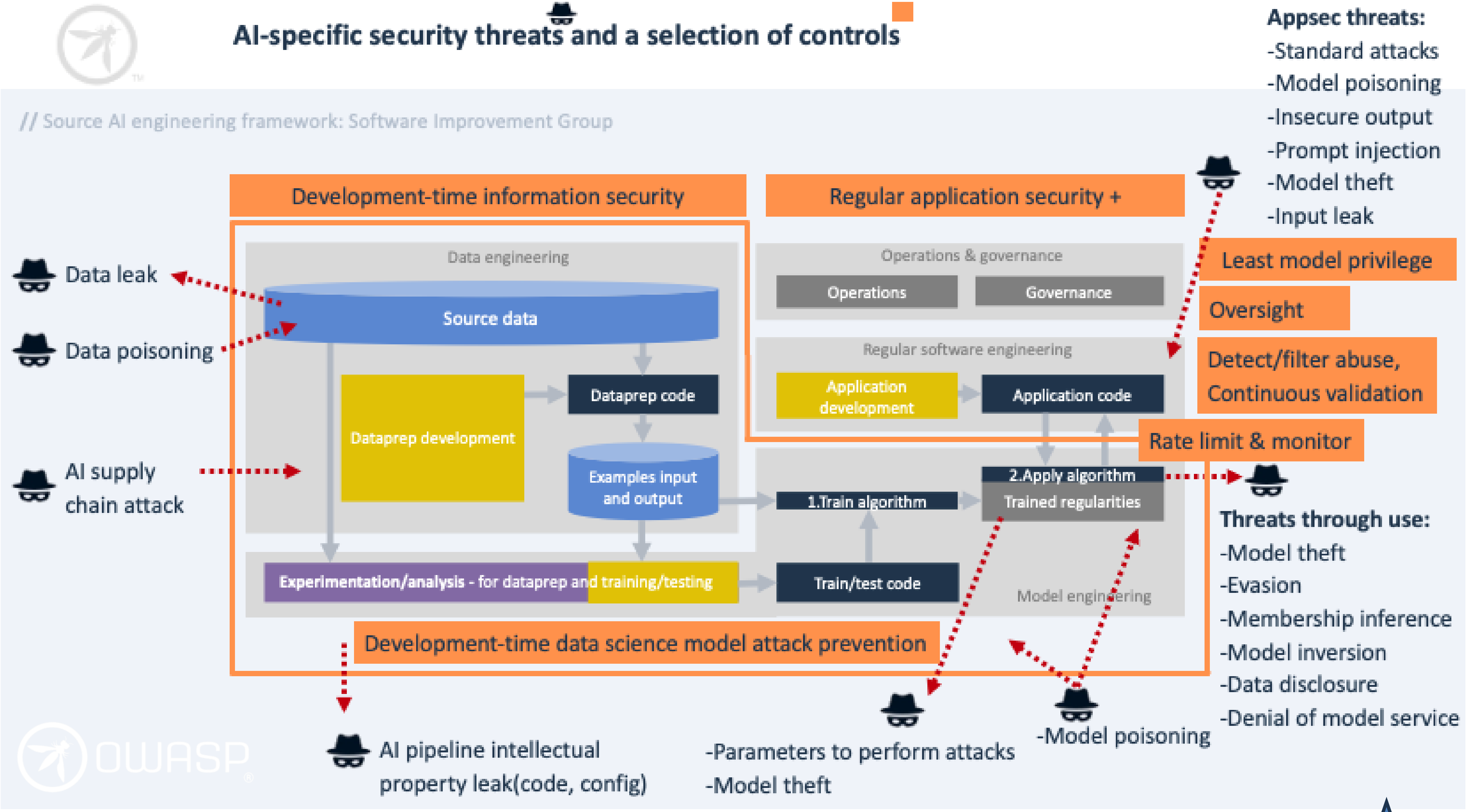
https://llmtop10.com/



1. ML01:2023 Input Manipulation Attack
2. ML02:2023 Data Poisoning Attack
3. ML03:2023 Model Inversion Attack
4. ML04:2023 Membership Inference Attack
5. ML05:2023 Model Theft
6. ML06:2023 AI Supply Chain Attacks
7. ML07:2023 Transfer Learning Attack
8. ML08:2023 Model Skewing
9. ML09:2023 Output Integrity Attack
10. ML10:2023 Model Poisoning

https://mltop10.info/

AI-specific security threats and a selection of controls



https://owaspai.org/

AI Security Matrix – OWASP AI Exchange

The AI security matrix below shows all threats and risks, ordered by attack surface and lifecycle.

AI-specific?	Lifecycle	Attack surface	Threat	Asset	Impacted	Unwanted result
AI	Runtime	Model use (provide input/ read output)	Direct prompt injection	Model behaviour	Integrity	Manipulated unwanted model behaviour causes wrong decisions leading to business financial loss, misbehaviour going undetected, reputational damage, legal and compliance issues, operational disruption, customer dissatisfaction and churn, reduced empoloyee morale, incorrect strastegic decisions, liabiity issues, personal damage and safety issues
			Indirect prompt injection			
			Evasion (e.g. adversarial examples)			
		Break into deployed model	Runtime model poisoning (reprogramming)			
	Development	Engineering environment	Development time model poisoning			
			Data poisoning of train/finetune data			
		Supply chain	Obtain poisoned foundation model (transfer learning attack)			
			Obtain poisoned data to train/finetune			
	Runtime	Model use	Unwanted disclosure in model output	Train data	Confidentiality	Leaking sensitive data can cause costs from fines and legal fees and remediation effort, loss of business through customer churn, reputation damage, loss of competitive advantage in case of trade secrets, operational disruption, impacted business relationships, and employee morale
			Model inversion / Membership inference			
	Development	Engineering environment	Train data leaks			
	Runtime	Model use	Model theft through by use (input-output harvesting)	Model intellectual property	Confidentiality	If attackers can copy a model, the investment in the model is devalued caused by loss of competitive advantage, plus a copy can help craft (evasion) attacks
		Break into deployed model	Runtime model theft			
	Development	Engineering environment	Development time model parameter leak			
	Runtime	Model use	System failure by use (model resource depletion)	Model behaviour	Availability	The model is not available, leading to business continuity issues, or safety problems
	Runtime	All IT	Model input leak	Model input data	Confidentiality	Sensitive data in model input leaks. E.g. an LLM prompt with a sensitive question, enhanced with retrieved company secrets
	Runtime	All IT	Model output contains injection attack	Any asset	C, I, A	Injection attack (from model output) causes harm
Generic	Runtime	All IT	Generic runtime security attack	Any asset	C, I, A	Generic runtime security attack causes harm (includes social engineering/phishing)
	Development	All IT	Generic supply chain attack	Any asset	C, I, A	Generic supply chain security attack causes harm (e.g. vulnerability in a component)

Source: OWASP AI Exchange at owaspai.org

https://owaspai.org/docs/ai_security_overview/

What You Can (Should) Do

Only You can Protect Yourself and Others

Secure By Design/Default

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

National Cyber Security Centre

National Cyber Security Centre

OAS

NISC

CSA SINGAPORE

certnz

NSM NORWEGIAN NATIONAL CYBER SECURITY CENTRE

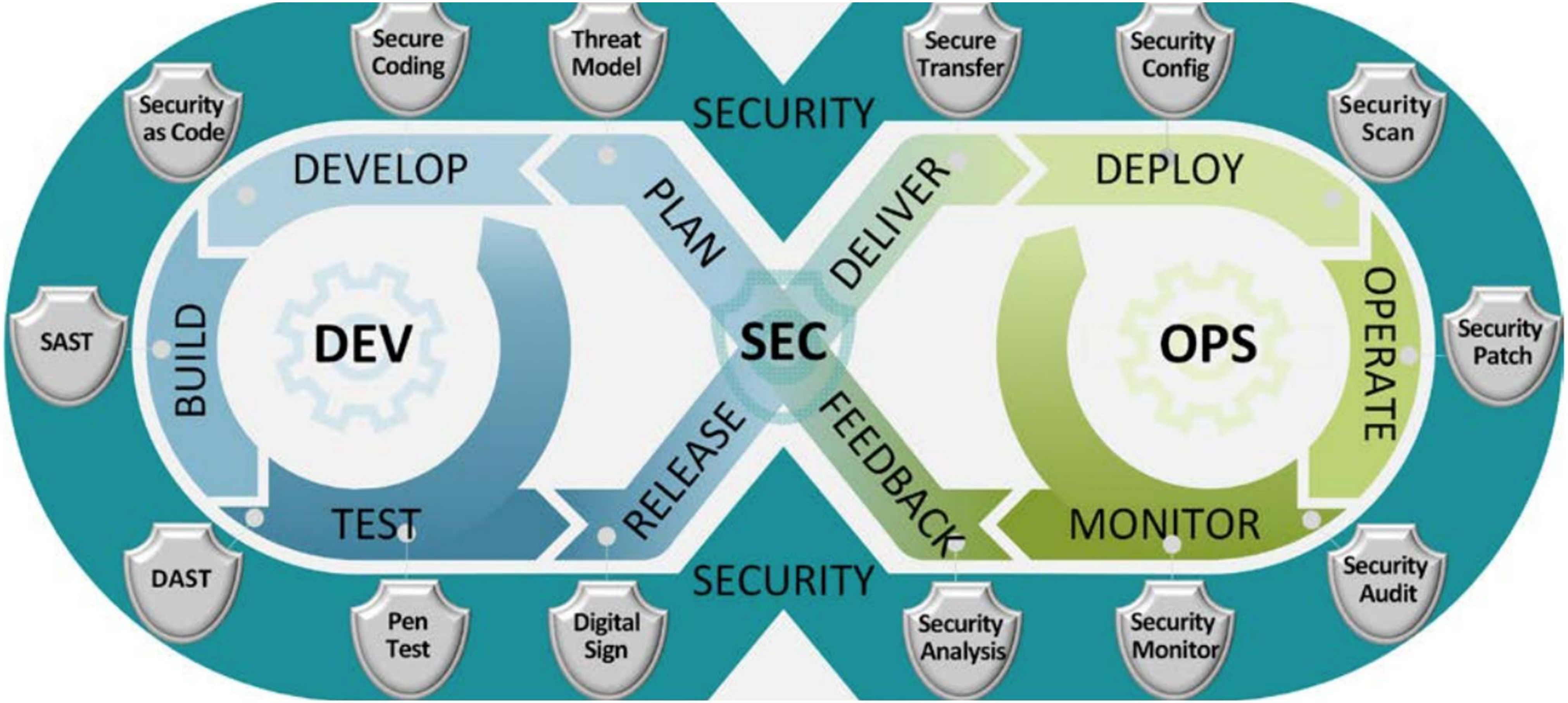
National Cyber and Information Security Agency

KISA KOREA INTERNET & SECURITY AGENCY

JPCERT CC®

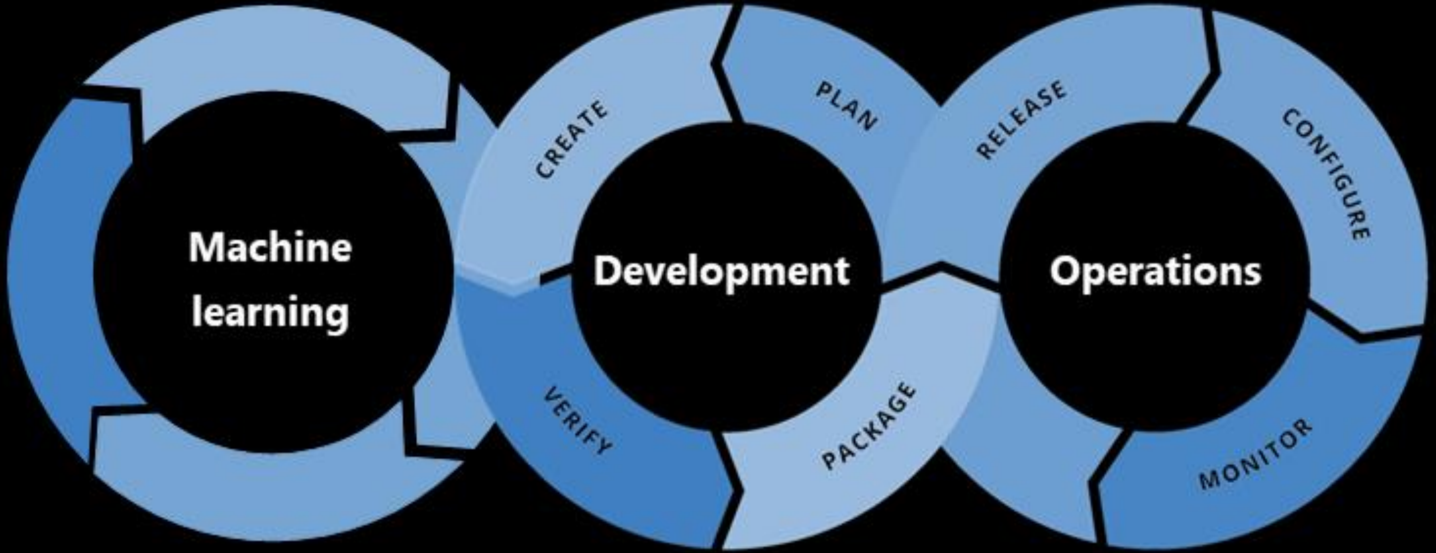
https://www.cisa.gov/securebydesign

Build Security In



Stages of DevSecOps Pipelines (from [DoD Enterprise DevSecOps Reference Design, Version 1.0](), 12 August 2019)

Machine learning operations:
Machine learning plus development and operations



Experiment
- Data acquisition
- Business understanding
- Initial modeling

Develop
- Modeling and testing
- Continuous integration
- Continuous deployment

Operate
- Continuous delivery
- Data feedback loop
- System and model monitoring

https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/how-to-approach-mlops

MLSecOps

TL;DR MLSecOps, please

MLSecOps stands for "Machine Learning Security Operations"

MLSecOps is the next evolution of building AI apps.

MLSecOps places security at the core of the ML Life Cycle.

MLSecOps merges traditional security and new concepts.

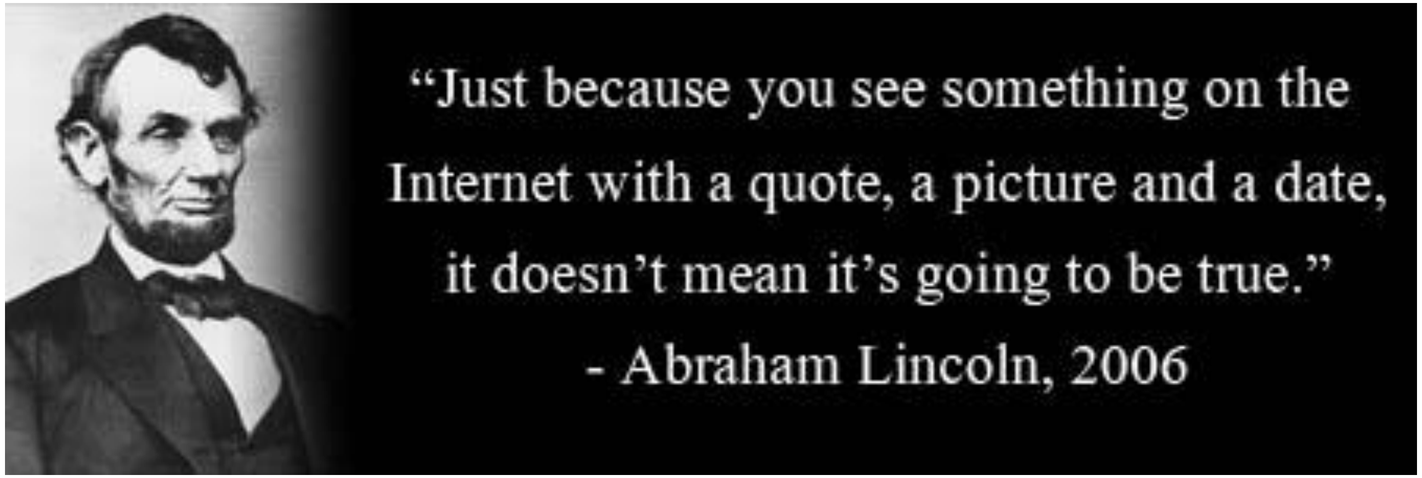
https://mlsecops.com/
https://mlsecops.com/what-is-mlsecops

Age of Zero Trust

Minimized footprint.

Assume breach.

Never trust. Always verify.



OWASP - Addressing AI Security

https://owaspai.org/docs/ai_security_overview/

- Implement **AI governance**
- **Extend security and development practices** to include data science activities especially to protect and streamline the engineering environment.
- **Improve regular application and system security** through understanding of AI particularities e.g. model parameters need protection and access to the model needs to be monitored and rate-limited.
- **Limit the impact** of AI by minimizing privileges and adding oversight, e.g. guardrails, human oversight.
- **Countermeasures in data science** through understanding of model attacks, e.g. data quality assurance, larger training sets, detecting common perturbation attacks, input filtering.

Guidelines for secure AI system development

National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

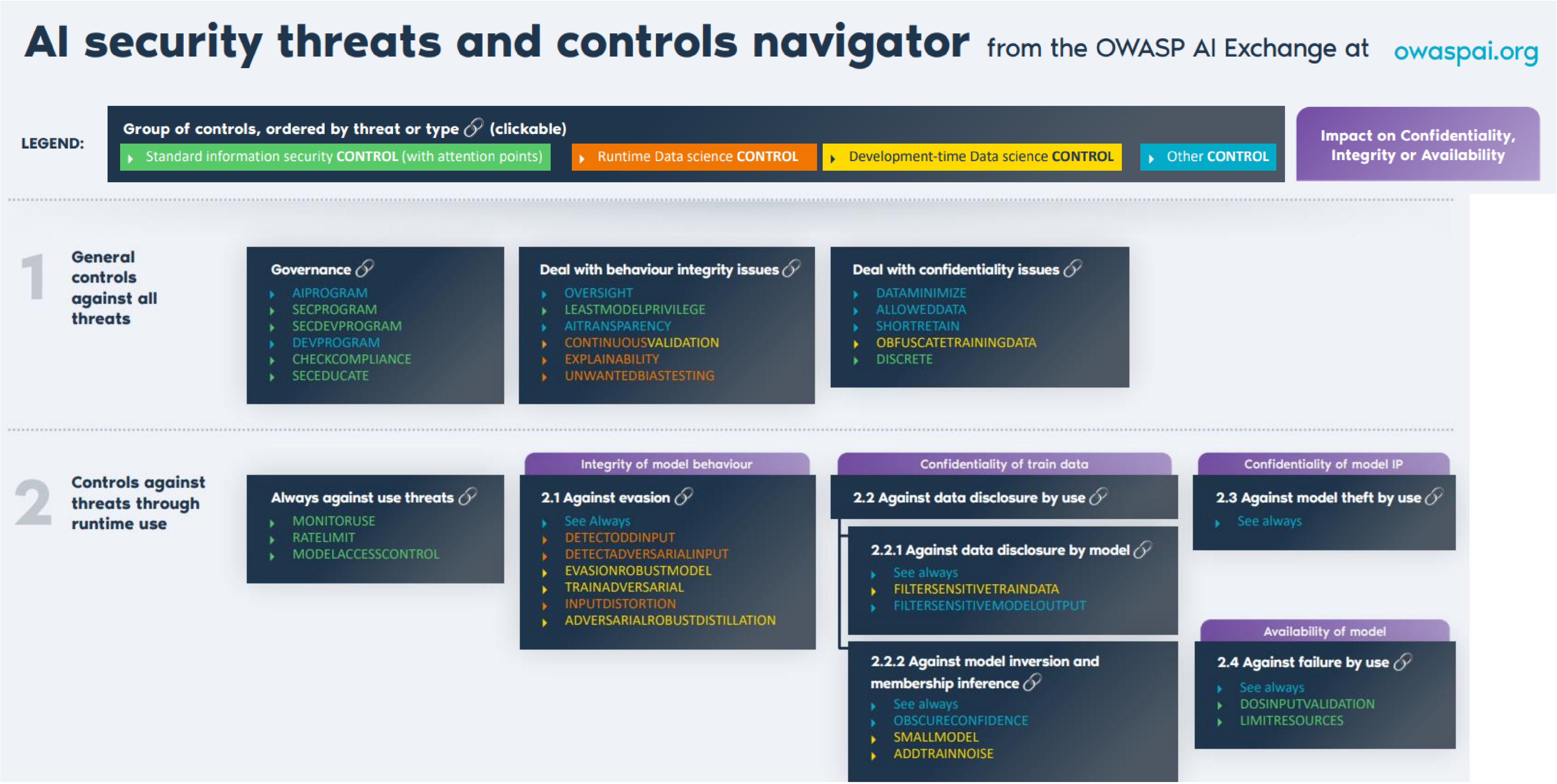
Guidelines for Secure AI Development,
https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf,
NCSC (UK) & CISA (US)

OWASP AI Exchange

Our Content

- AI Security Overview
- 1. General controls
- 2. Threats through use
- 3. Development-time threats
- 4. Runtime application security threats

https://owaspai.org/



https://github.com/OWASP/www-project-ai-security-and-privacy-guide/raw/main/assets/images/owaspaioverviewpdfv3.pdf

2023 2024 CISA ROADMAP FOR ARTIFICIAL INTELLIGENCE

https://www.cisa.gov/resources-tools/resources/roadmap-ai

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

Trustworthy & Responsible AI Resource Center

https://airc.nist.gov/Home

AI Risk Management Framework (RMF)

The AI RMF is voluntary guidance to improve the ability to incorporate trustworthiness considerations into the design, development, use and evaluation of AI products, services and systems.

Download the Framework

AI RMF Playbook

Companion resource for the AI RMF that includes suggested actions, references, and documentation guidance to achieve outcomes for the four AI RMF functions.

Download the Playbook (as PDF) (as CSV) (as JSON)

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

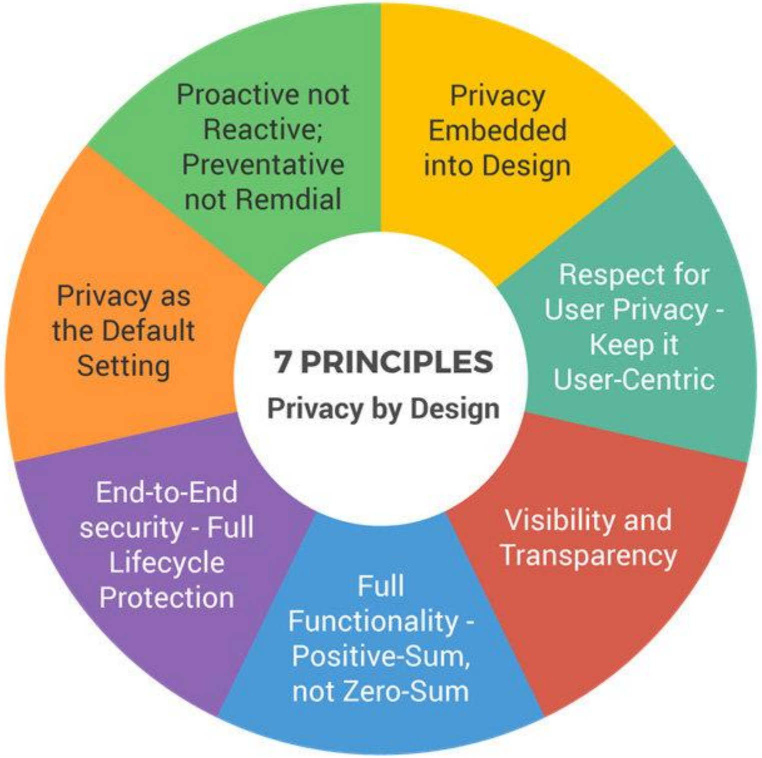
Trustworthy & Responsible AI Resource Center

NIST AI Risk Management Framework

AI Risks & Trustworthiness

1. Valid and Reliable
2. Safe
3. Secure and Resilient
4. Accountable and Transparent
5. Explainable and Interpretable
6. Privacy-Enhanced
7. Fair – with Harmful Bias Managed

https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf



Cavoukian, A., ***Privacy by Design, The 7 Foundational Principles, Implementation and Mapping of Fair Information Practices***, https://privacy.ucsc.edu/resources/privacy-by-design---foundational-principles.pdf

AI Use in Cybersecurity

Prompt: Provide 5 ways AI can help cybersecurity and technical professionals reduce risks of malicious use of AI.

Threat Intelligence and Prediction:

- ***Problem***: Staying ahead of evolving attacker tactics and emerging threats is a constant challenge for security teams.
- ***Solution***: -powered threat intelligence platforms aggregate data from various sources, analyze attack trends, and predict future threats. This proactive approach allows security teams to prioritize defenses, allocate resources effectively, and anticipate new attack vectors before they materialize.

Automated Threat Detection and Response:

- ***Problem***: Traditional security approaches rely on manual analysis of alerts and incident response, causing delays and potential breaches.
- ***Solution***: AI-powered tools analyze network traffic, system logs, and user behavior in real-time, identifying anomalies and triggering automated responses like quarantining suspicious files, blocking malicious IP addresses, or notifying security teams for immediate action. This significantly reduces response times and minimizes damage from potential attacks.

Phishing and Social Engineering Detection:

- ***Problem***: Sophisticated phishing attacks and social engineering tactics can bypass traditional filters and deceive users.
- ***Solution***: AI-powered analysis can assess the linguistic style, content, and sender information of emails to identify suspicious patterns and predict phishing attempts. Additionally, AI can analyze chatbots and user interactions to detect suspicious behaviors and prevent social engineering attacks.

AI Use in Cybersecurity

Prompt: Provide 5 ways AI can help cybersecurity and technical professionals reduce risks of malicious use of AI.

Proactive Vulnerability Management:

- ***Problem***: Unpatched vulnerabilities and misconfigurations offer easy entry points for attackers.
- ***Solution***: AI-powered scanning tools can analyze vast codebases and system configurations to identify vulnerabilities and misconfigurations before they're exploited. This proactive approach prioritizes critical vulnerabilities and automates patch deployment, minimizing attack surfaces and strengthening overall security posture.

Endpoint Security with Behavioral Analysis:

- ***Problem***: Traditional endpoint security relies on signature-based detection, missing zero-day attacks and other novel threats.
- ***Solution***: AI-powered endpoint protection establishes baselines of normal endpoint behavior and continuously monitors deviations. This allows for real-time anomaly detection, even for unknown threats, and targeted interventions to prevent malware execution and data breaches.

Questions

Hacking AI – Ron Woerner – Feb 1, 2024

“Apply” / Next Steps

In the next week:

- ➢ Review the slide deck
- ➢ Pick 2-3 references for further learning

In the next month:

- ➢ Review the slide deck
- ➢ Pick 2-3 other references for further learning
- ➢ Teach at least 1 other person what you’ve learned

Slides available at https://github.com/hackerron/AI-Cybersecurity

Resources & References

https://github.com/hackerron/AI-Cybersecurity

- Kumar, et.al., *Failure modes in machine learning*. Microsoft Learn, Microsoft, (2022, November 2). https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning
- NCSC (UK) & CISA (US), *Guidelines for Secure AI Development*, https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf
- NCSC, *The near-term impact of AI on the cyber threat,* January 24, 2024, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat
- CISA, *2023-2024 Roadmap for Artificial Intelligence*, November 2023, https://www.cisa.gov/sites/default/files/2023-11/2023-2024_CISA-Roadmap-for-AI_508c.pdf
- NIST AI Risk Management Framework 1.0, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
- Cloud Security Alliance (CSA), *AI Safety Initiative*, https://cloudsecurityalliance.org/research/working-groups/artificial-intelligence/
  CSA, *Security Implications of ChatGPT*, Aug 2, 2023, https://cloudsecurityalliance.org/artifacts/security-implications-of-chatgpt/
- MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems), https://atlas.mitre.org/
- OWASP AI Exchange, https://owaspai.org/
  - OWASP Top 10 for LLM Applications, https://llmtop10.com/
  - OWASP Machine Learning Top 10 (2023 ed draft), https://mltop10.info/
  - OWASP AI Security Matrix, https://owaspai.org/docs/ai_security_overview/
  - OWASP Project AI Security and Privacy Guide, https://github.com/OWASP/www-project-ai-security-and-privacy-guide/blob/main/owaspaiexchange.md

Other Resources

Training course:

- Kelly, D., *Security Risks in AI and Machine Learning: Categorizing Attacks and Failure Modes*, LinkedIn Learning, Feb 23, 2022, https://www.linkedin.com/learning/security-risks-in-ai-and-machine-learning-categorizing-attacks-and-failure-modes/

Books:

- Hutchens, J., *The Language of Deception: Weaponizing Next Generation AI*, (2023), Wiley, ISBN-13: 978-1394222544, https://www.amazon.com/Language-Deception-Weaponizing-Next-Generation/dp/1394222548
- Baker, P., *ChatGPT for Dummies*, (2023), ISBN-13: 978-1394204632

Slides available at https://github.com/hackerron/AI-Cybersecurity

Hacking AI
Risks and Rewards For Cybersecurity

AI Omaha Meetup
Feb 1, 2024
Ron Woerner

LinkedIn:
https://www.linkedin.com/in/ronwoerner/

Slides available at https://github.com/hackerron/AI-Cybersecurity