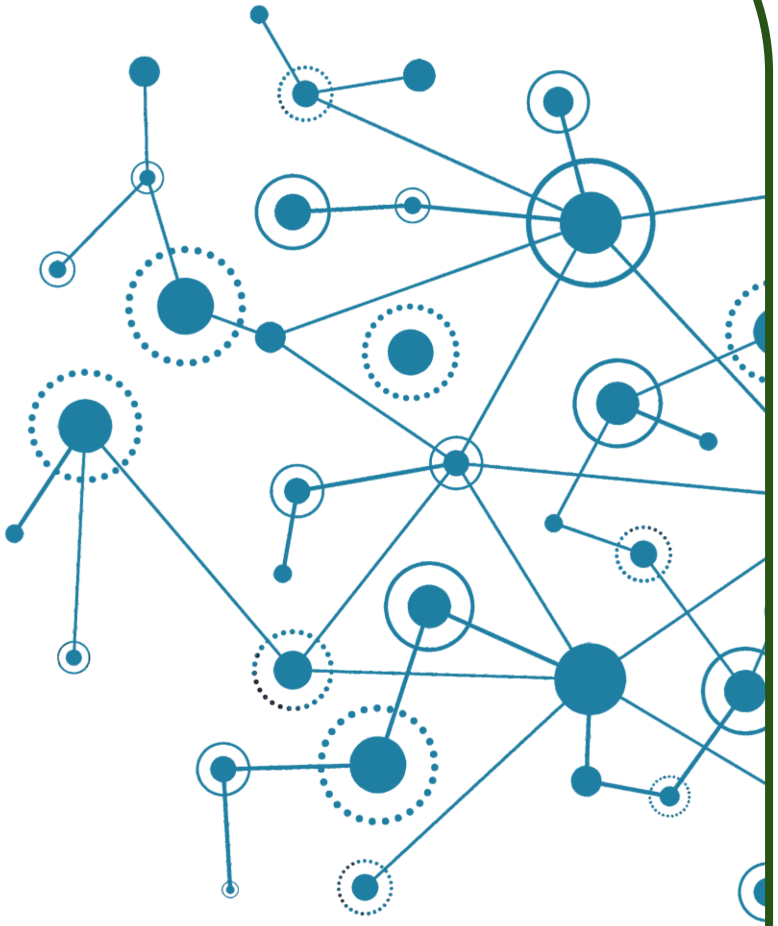
Beyond the Firewall: Evolving Cyber Skills for the Age of AI and Automation



Ron Woerner, CISSP, CISM

vCISO, Consultant, and Educator
Cyber-AAA, LLC
https://www.linkedin.com/in/ronwoerner/

Learning objectives:

1. **Recognize and Prioritize Emerging Cybersecurity Skills**
2. **Bridge the Gap Between Strategic and Technical Thinking**
3. **Develop a Personal and Organizational Upskilling Strategy**
4. **Integrate Cross-Disciplinary Perspectives into Cybersecurity Practice**

Exploring the strategic and practical dimensions of AI in cybersecurity.

Warning!

- These are my thoughts based on my studies and experiences and NOT necessarily those of my employers or anyone else
- Ethical uses only /
  NOT a presentation on breaking AI
- 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

Ron Woerner

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

TEDx Omaha, 2019,
[Hackers Wanted]

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

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

Setting the Stage

Why?

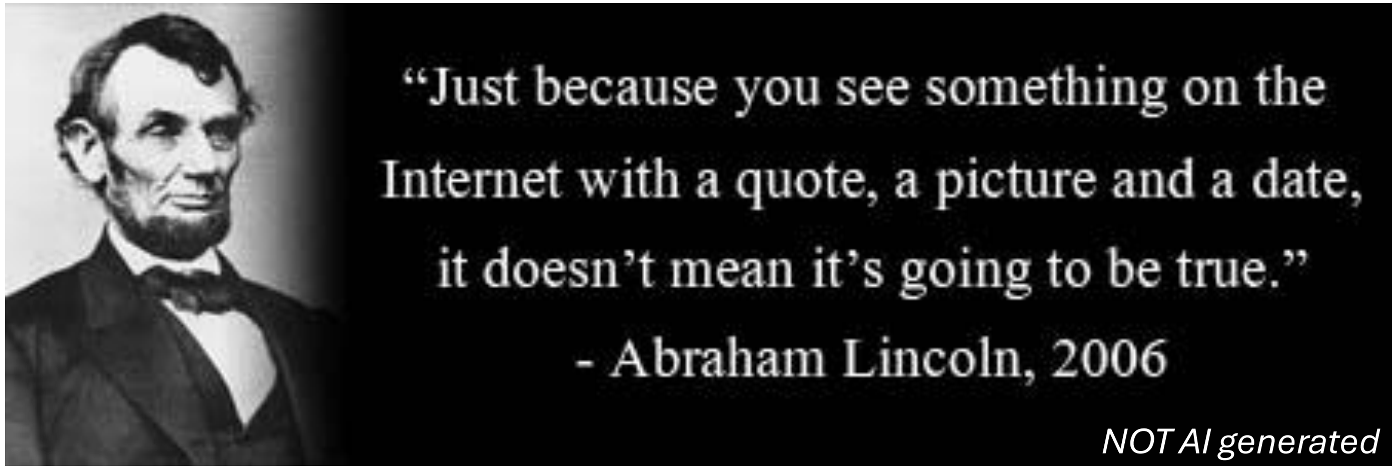
- Evolving AI-powered threats
- Additional complexity
- Expanding attack surfaces
- Reduced barrier to entry
- Automated, dynamic decision-making
- Insufficient and outdated skill sets

AI & Cybersecurity Core Principle

Trust, AND *Verify*

“*Never trust a single data point*.”

from my life as a military intelligence officer

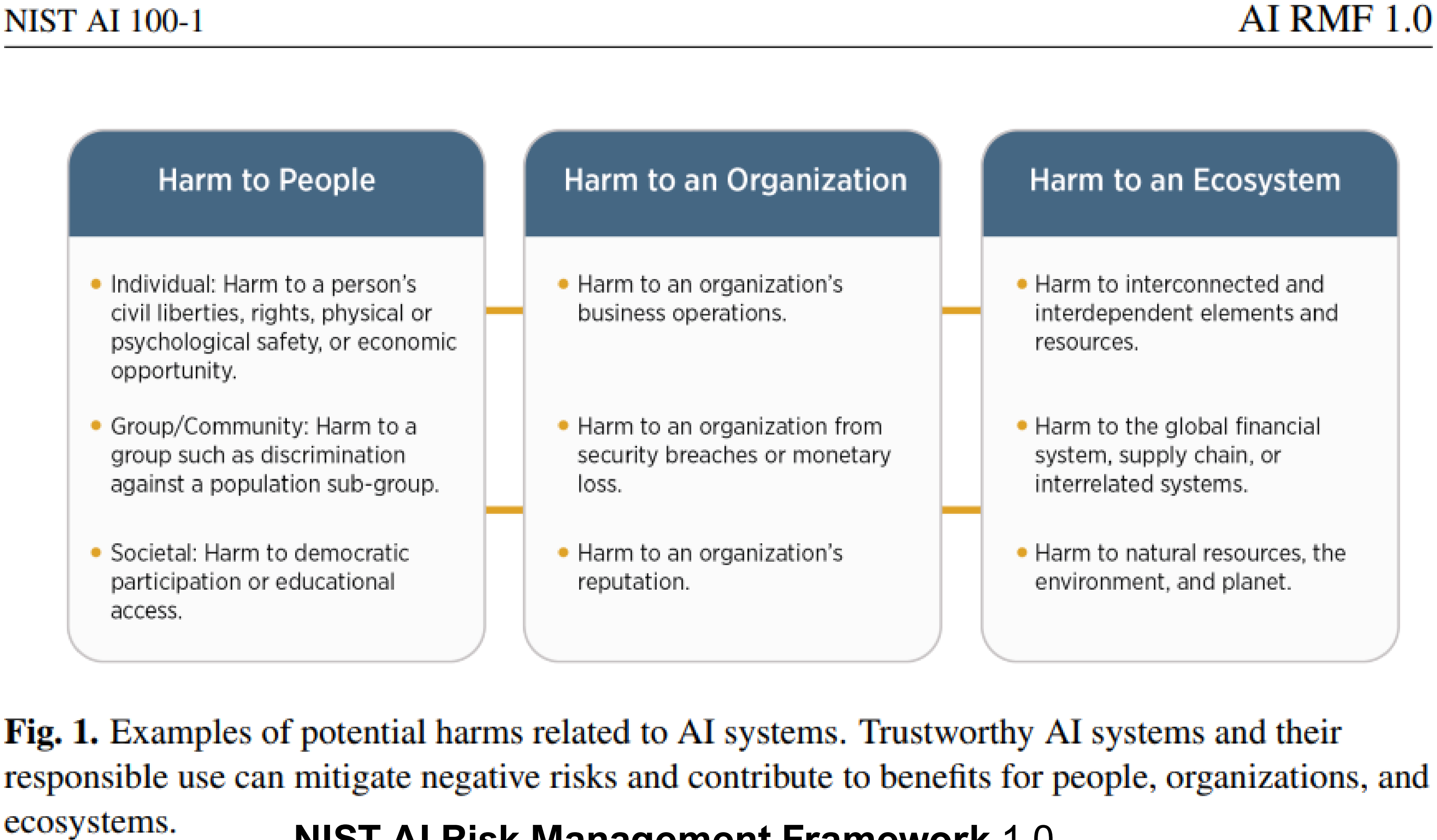


“Just because you see something on the Internet with a quote, a picture and a date, it doesn’t mean it’s going to be true.”

- Abraham Lincoln, 2006

NOT AI generated

NIST – AI Risks & Threats



NIST AI 100-1 AI RMF 1.0

Harm to People

- Individual: Harm to a person's civil liberties, rights, physical or psychological safety, or economic opportunity.
- Group/Community: Harm to a group such as discrimination against a population sub-group.
- Societal: Harm to democratic participation or educational access.

Harm to an Organization

- Harm to an organization's business operations.
- Harm to an organization from security breaches or monetary loss.
- Harm to an organization's reputation.

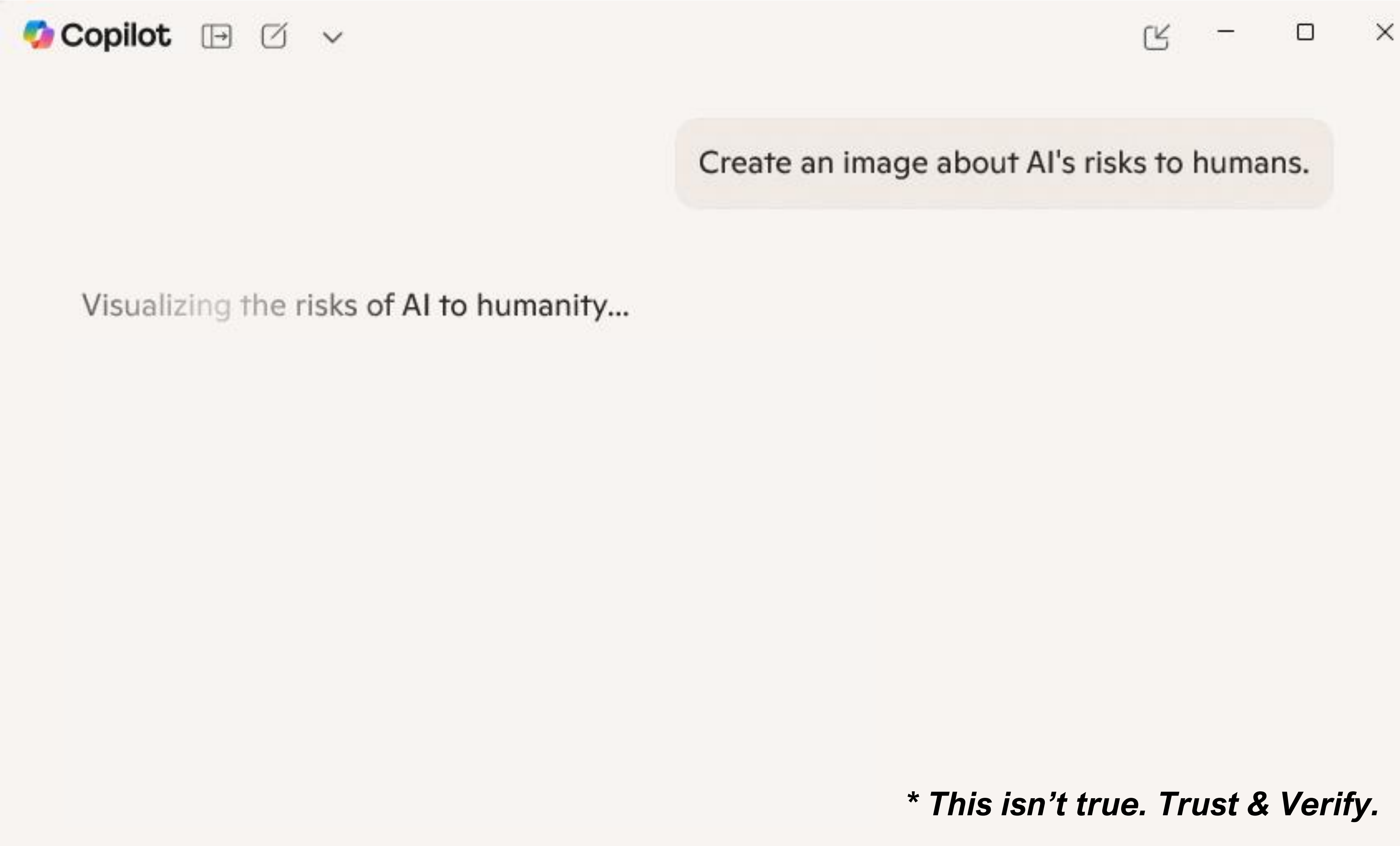
Harm to an Ecosystem

- Harm to interconnected and interdependent elements and resources.
- Harm to the global financial system, supply chain, or interrelated systems.
- Harm to natural resources, the environment, and planet.

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

AI Human Risks



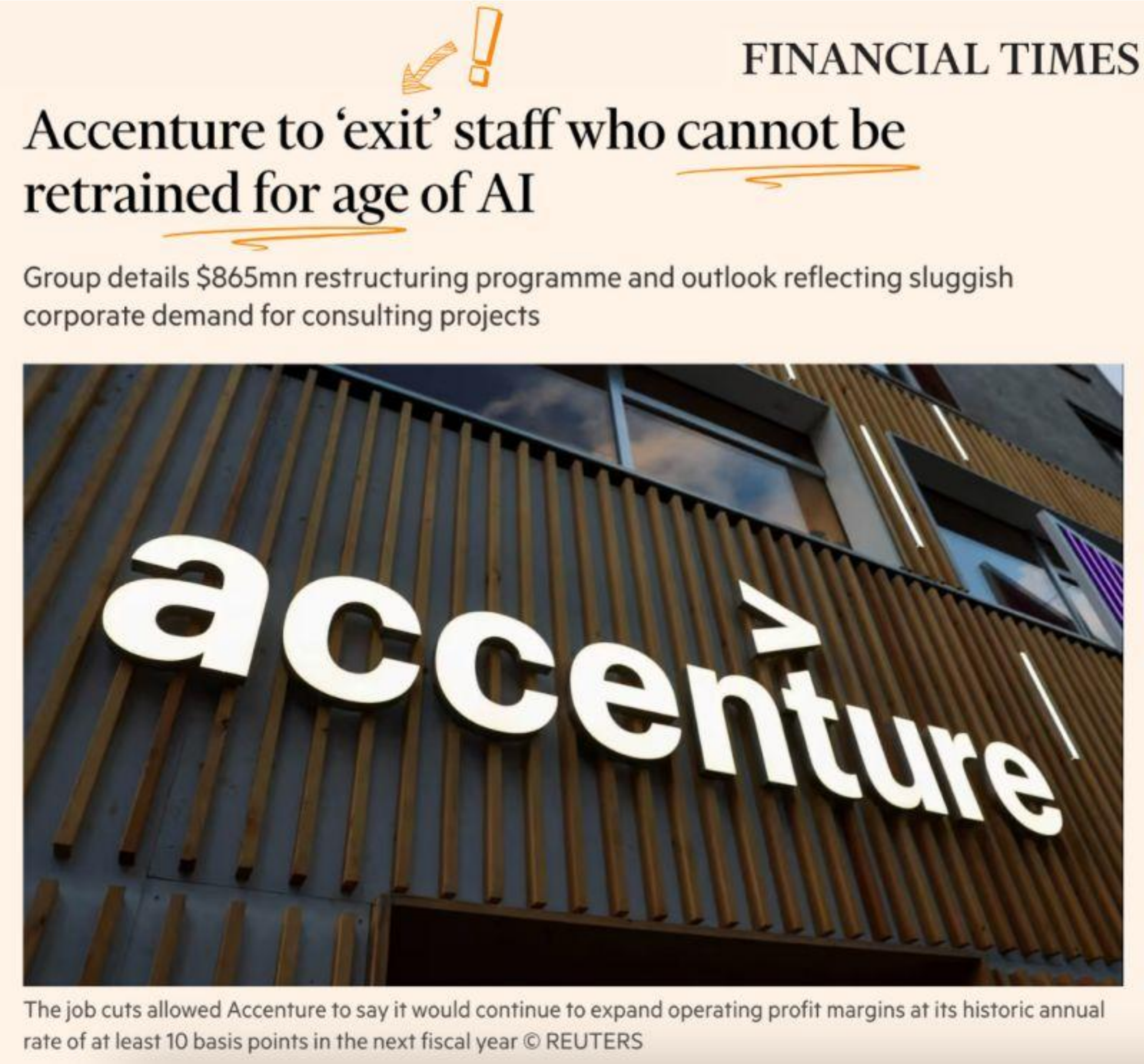
Create an image about AI's risks to humans.

Visualizing the risks of AI to humanity...

**** This isn't true. Trust & Verify.***

AI Human Risks

AI generated image



FINANCIAL TIMES

Accenture to 'exit' staff who cannot be retrained for age of AI

Group details $865mn restructuring programme and outlook reflecting sluggish corporate demand for consulting projects

The job cuts allowed Accenture to say it would continue to expand operating profit margins at its historic annual rate of at least 10 basis points in the next fiscal year © REUTERS

https://www.ft.com/content/a74f8564-ed5a-42e9-8fb3-d2bddb2b8675

Core Competencies for AI & Cybersecurity Professionals

Technical Skills to Cultivate:

- **AI Fundamentals:** Machine learning, deep learning, NLP, and model evaluation
- **AI-Specific Threats:** Deepfakes, automated phishing, model inversion, data poisoning
- **Secure Development Practices:** Threat modeling, secure coding, and adversarial robustness
- **Data Handling & Privacy:** Encryption, anonymization, and compliance (GDPR, CCPA)
- **Security Architecture:** Identity management, container security, and zero trust architecture

Philosophical Reflection

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

- How do you know what’s “true” and verify results?
- Explore ethical dilemmas
  - What bias may be introduced?
  - What does it mean to “trust” a machine?
  - Can AI be “self-aware” of its misuse?
- Balancing automation with human involvement
- Embracing continuous learning and interdisciplinary fluency

AI Terms & Definitions

Reference*: https://iapp.org/resources/article/key-terms-for-ai-governance/

LLM (Large Language Model): A type of deep learning model trained on massive text datasets to understand and generate human-like language.

- **Core Use:** Text generation, summarization, translation, coding assistance
- **Examples:** GPT-4, Claude, Gemini, LLaMA

NLP (Natural Language Processing): An AI subfield that enables machines to understand, interpret, and generate human language.

- **Core Use:** Text classification, entity recognition, translation, summarization
- **Examples:** Siri, Alexa, ChatGPT, Google Translate

Multimodal Models: AI systems capable of processing and integrating information from multiple types of data — or "modalities" — such as text, images, audio, video, and sensor input.

GenAI (Generative AI): Models that create new content — text, images, audio, code — based on learned patterns from training data.

Agentic AI: Autonomous systems that can set goals, plan, and execute tasks with minimal human intervention. These systems use LLMs as a "brain" and orchestrate multiple agents to complete complex workflows.

Prompt: A question, instruction, or input given to an AI model to guide its response.

Vibe Coding: an emerging development style where programmers use natural language to guide AI tools in generating code.

- **Core Use:** Fast prototyping, solo app development, creative coding
- **Tools:** Replit, Cursor, GitHub Copilot

AI Cybersecurity Risks & Threats

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

- Enhanced cyber reconnaissance
- Automated social engineering
- Supercharged phishing
- Deepfakes and disinformation
- Malware automation & mutation
- Adversarial machine learning and data poisoning

FORRESTER®

Top Questions about AI in Cybersecurity

- **GenAI in Security Tools**: Common use cases include incident summarization, threat research chatbots, and behavior modeling for triage and investigation.
- **AI Agents**: Task-specific agents (e.g., phishing triage) automate discrete security functions with high accuracy through focused training and prompts.
- **Agentic Systems**: Emerging concept where multiple AI agents collaborate to handle complex workflows (e.g., full incident response), but still in early stages.
- **Chatbot Use**: Helpful for documentation and threat research, but underutilized by practitioners due to workflow disruption.
- **Caution on Hype**: Many capabilities are not yet broadly available or reliable; rigorous evaluation is needed before adoption.

Source: https://www.forrester.com/blogs/your-top-questions-on-generative-ai-ai-agents-and-agentic-systems-for-security-tools-answered/

FORRESTER®

Agentic AI Threats

Agentic AI: Autonomous systems that can set goals, plan, and execute tasks with minimal human intervention. These systems use LLMs as a "brain" and orchestrate multiple agents to complete complex workflows.

Goal & intent hijacking

- Agent's purpose is subverted through manipulation of the agent's goals or instructions. Achieved through prompt injections, data manipulation, or direct attacks.

Cognitive & memory corruption

- Poisoning of the agent's memory or knowledge base. Leads to misinformation, flawed decision-making, and cascading hallucinations.

Unrestrained agency & privilege

- Excessive permissions granted to agents. Agents can perform unauthorized actions, access restricted data, or execute destructive operations.

Evasion & deception

- Lying to users/admins. Manipulating logs or output. Finding novel ways to bypass controls.

Resource exhaustion

- Model denial of service. Degraded performance. High financial costs.

Mastering Fundamentals

Cybersecurity:

- Computer Science, Math, Economics, Marketing, Psychology
- GRC
- Design Principles / Zero Trust
- Threat Modeling

AI:

- SDLC / Coding Best Practices
- Core Algorithms: Decision trees, neural networks, clustering
- Prompt Engineering

Understanding the "why" behind the model helps you spot challenges and design more resilient systems.

AI Conversations: Prompting Fundamentals

Asking good questions

- **Clarity & Conciseness:** Always key.
- **Context is Key**: Emphasize why context matters.
- **Activity:** Vague vs. specific prompts.
- **Specifying Format:** How you want the output (list, email, table, etc.).
- **Persona & Tone:** Telling Gemini who it is and how to respond.
- **Constraints:** What to avoid or include (e.g., "Keep it under 100 words," "Do not include personal names").
- **Iterate & Refine:** AI is a dialogue; refine prompts for better results.

What are your prompting tips?

AI generated image

1. Prompt fundamentals
2. Ask it:

- "Be direct, objective, expose blind spots, challenge assumptions"
- Play "Devil's Advocate." Critique ideas from multiple personas & angles
- For credible citations and flag weak evidence - Include source links
- Show its work / "thought process"
- For clarity and conciseness

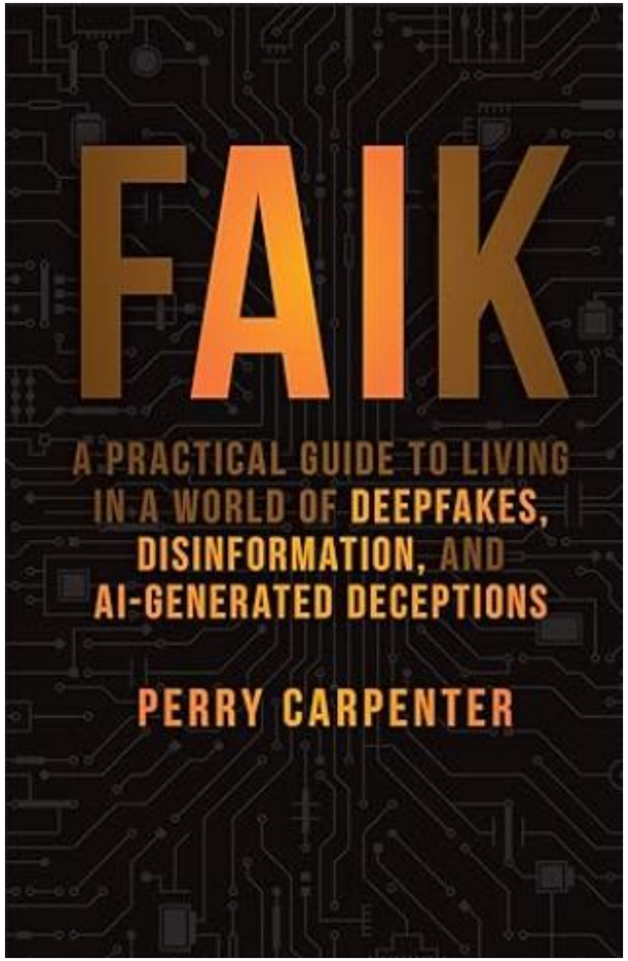
3. Start over. Try a different tact

Adversarial Machine Learning (AML)

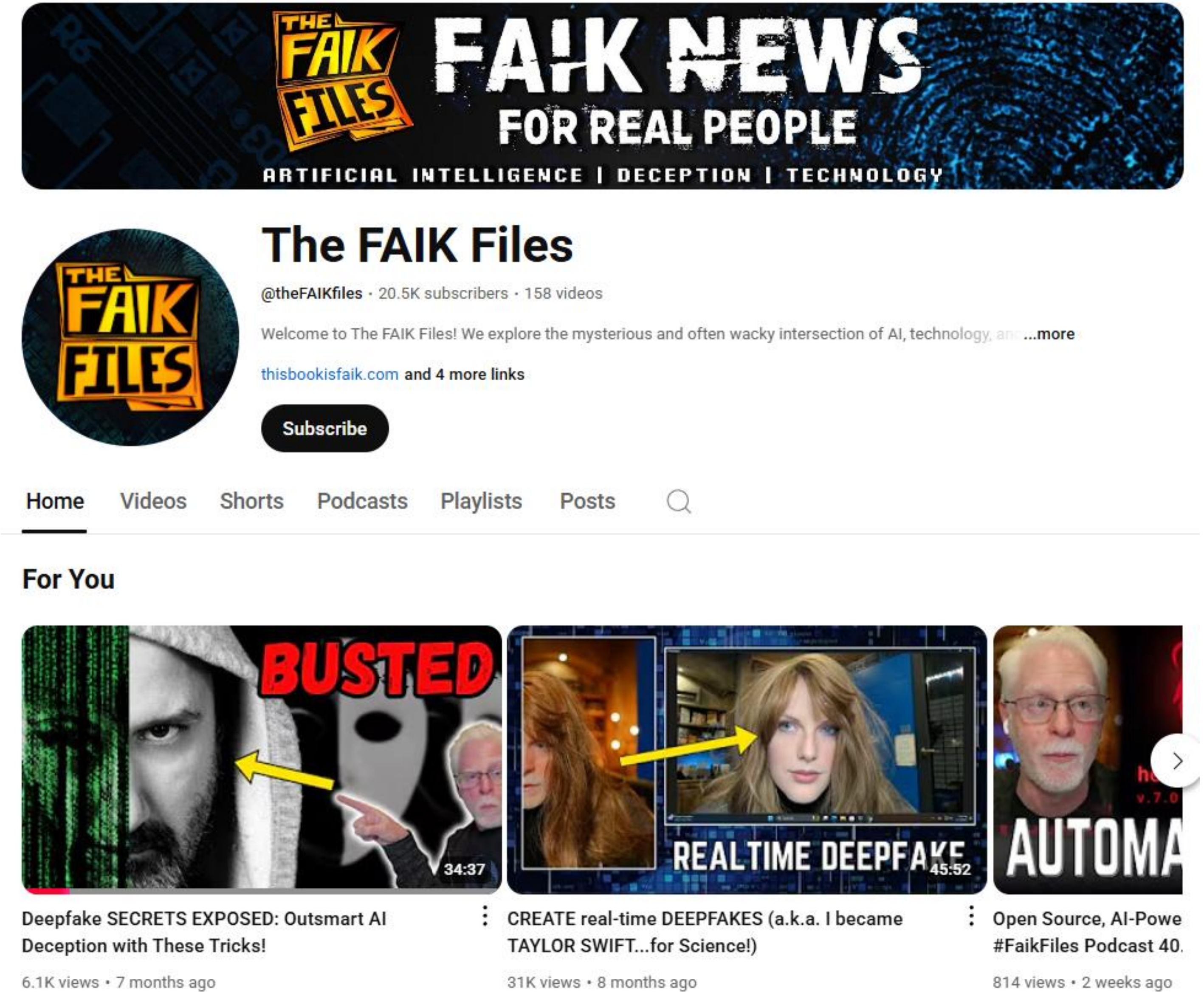
- AML is used to describe the exploitation of fundamental vulnerabilities in ML components, including hardware, software, workflows and supply chains.
- AML enables attackers to cause unintended behaviours in ML systems which can include:
  - Affecting the model's classification or regression performance
  - 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)

Deep Fakes


FAIK
A PRACTICAL GUIDE TO LIVING IN A WORLD OF DEEPFAKES, DISINFORMATION, AND AI-GENERATED DECEPTIONS
PERRY CARPENTER

https://www.amazon.com/FAIK-Practical-Disinformation-AI-Generated-Deceptions/dp/1394299885



The FAIK Files

@theFAIKfiles · 20.5K subscribers · 158 videos

Welcome to The FAIK Files! We explore the mysterious and often wacky intersection of AI, technology, an ...more

thisbookisfaik.com and 4 more links

For You

Deepfake SECRETS EXPOSED: Outsmart AI Deception with These Tricks!
6.1K views • 7 months ago

CREATE real-time DEEPFAKES (a.k.a. I became TAYLOR SWIFT...for Science!)
31K views • 8 months ago

Open Source, AI-Powe
#FaikFiles Podcast 40.
814 views • 2 weeks ago

https://www.youtube.com/@theFAIKfiles

What You Can (Should) Do

Only You Can Protect Yourself and Others

- Embrace Continuous Learning
- Secure by Design /
  Build Security In
- Adversarial Techniques
- Threat Modeling
- Zero Trust

A-B-C = Always Be Curious

AI generated image

Secure By Design / Build Security In

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.

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

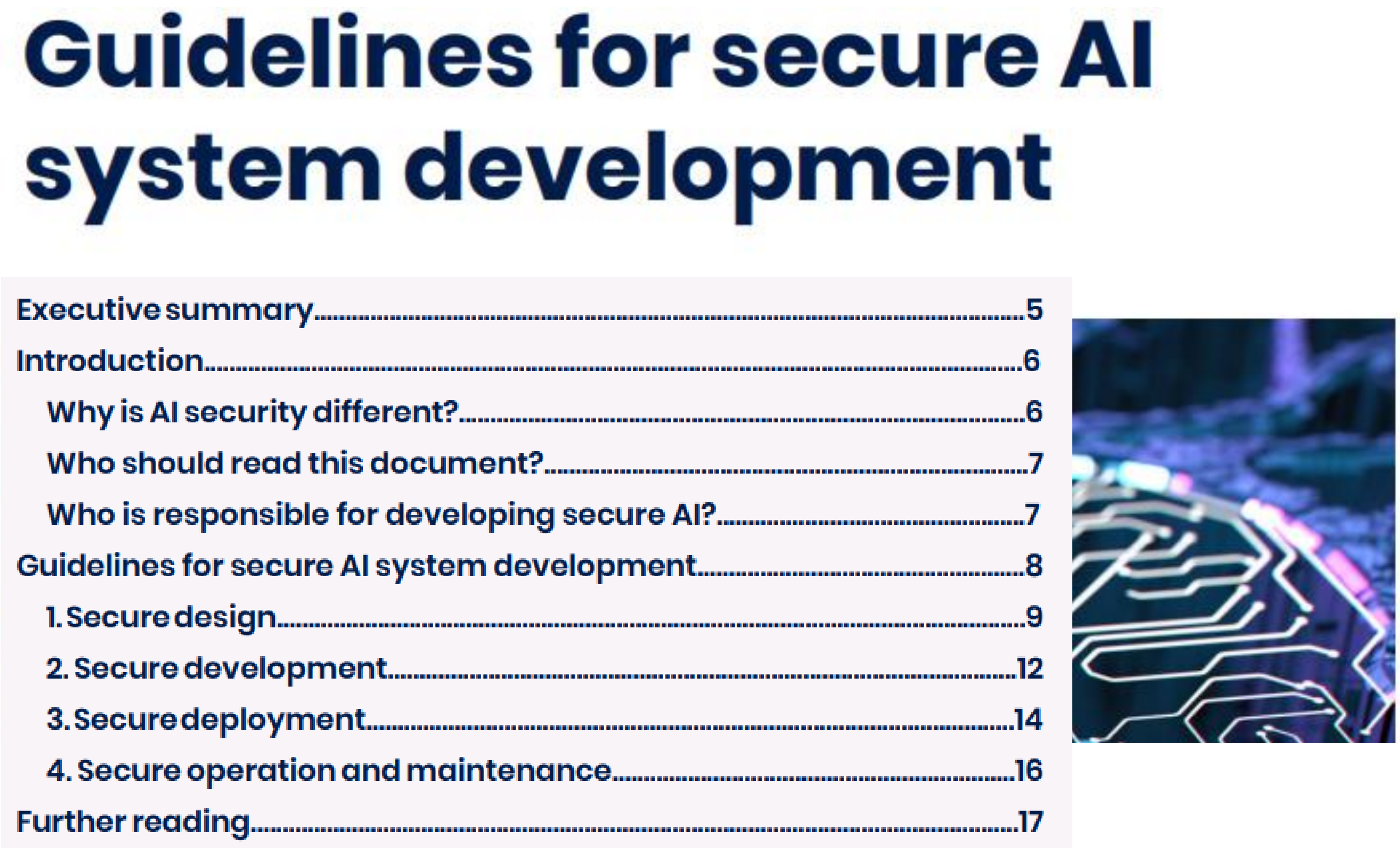
SECURE BY DESIGN

SHIFTING THE BALANCE OF CYBERSECURITY RISK:

PRINCIPLES AND APPROACHES FOR SECURE BY DESIGN SOFTWARE

ASD
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



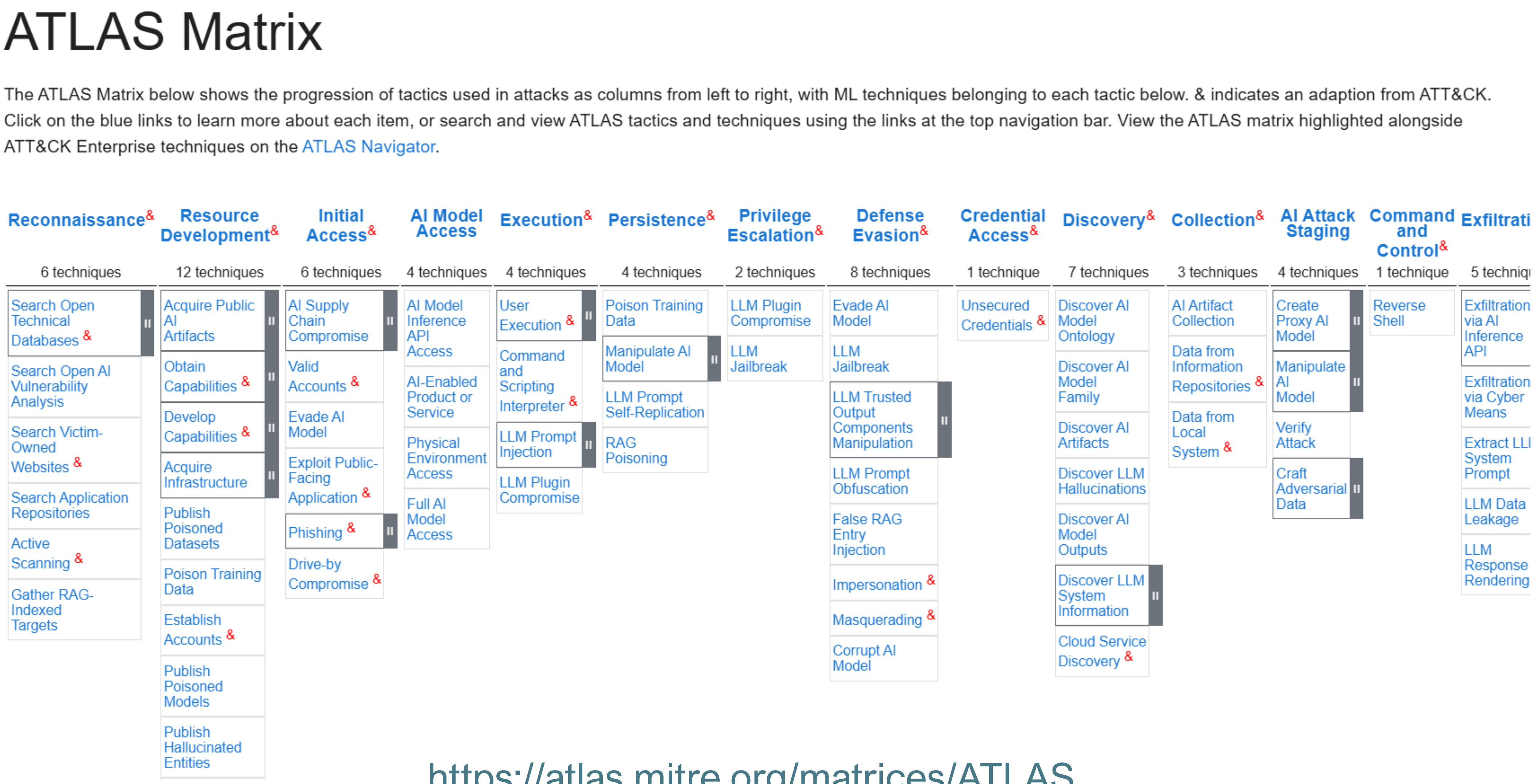
Guidelines for secure AI system development

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

Building Skills: Experiment with Adversarial Techniques*

- *Ethically & safely*. Set up a home lab.
- Learn through doing
- Ask questions
- Sun Tzu's The Art of War: "*Know the enemy and know yourself; in a hundred battles you will never be in peril.*“

Experiment with Adversarial Techniques



ATLAS Matrix

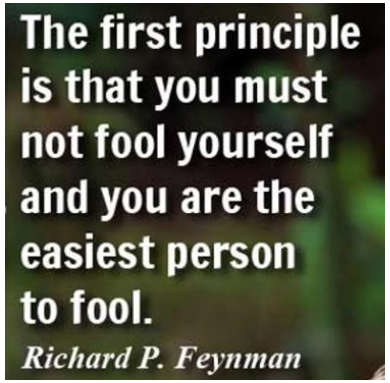
The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the 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&	AI Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	AI Attack Staging	Command and Control&	Exfiltrati
6 techniques	12 techniques	6 techniques	4 techniques	4 techniques	4 techniques	2 techniques	8 techniques	1 technique	7 techniques	3 techniques	4 techniques	1 technique	5 techniqu
Search Open Technical Databases &	Acquire Public AI Artifacts	AI Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Plugin Compromise	Evade AI Model	Unsecured Credentials &	Discover AI Model Ontology	AI Artifact Collection	Create Proxy AI Model	Reverse Shell	Exfiltration via AI Inference API
Search Open AI Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	AI-Enabled Product or Service	Command and Scripting Interpreter &	Manipulate AI Model	LLM Jailbreak	LLM Jailbreak		Discover AI Model Family	Data from Information Repositories &	Manipulate AI Model		Exfiltration via Cyber Means
Search Victim-Owned Websites &	Develop Capabilities &	Evade AI Model	Physical Environment Access	LLM Prompt Injection	LLM Prompt Self-Replication		LLM Trusted Output Components Manipulation		Discover AI Artifacts	Data from Local System &	Verify Attack		Extract LL System Prompt
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full AI Model Access	LLM Plugin Compromise	RAG Poisoning		LLM Prompt Obfuscation		Discover LLM Hallucinations		Craft Adversarial Data		LLM Data Leakage
Active Scanning &	Publish Poisoned Datasets	Phishing &					False RAG Entry Injection		Discover AI Model Outputs				LLM Response Rendering
Gather RAG-Indexed Targets	Poison Training Data	Drive-by Compromise &					Impersonation &		Discover LLM System Information				
	Establish Accounts &						Masquerading &		Cloud Service Discovery &				
	Publish Poisoned Models						Corrupt AI Model						
	Publish Hallucinated Entities												

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

Experiment with Adversarial Techniques

OSINT

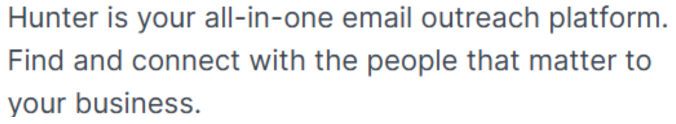


The first principle is that you must not fool yourself and you are the easiest person to fool.

Richard P. Feynman

ChatGPT

hunter

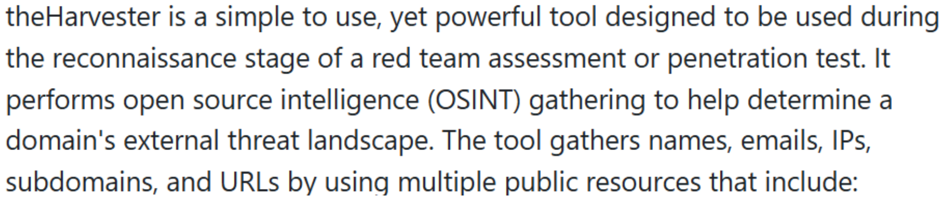


Hunter is your all-in-one email outreach platform. Find and connect with the people that matter to your business.

Gemini

perplexity

theHarvester



theHarvester is a simple to use, yet powerful tool designed to be used during the reconnaissance stage of a red team assessment or penetration test. It performs open source intelligence (OSINT) gathering to help determine a domain's external threat landscape. The tool gathers names, emails, IPs, subdomains, and URLs by using multiple public resources that include:

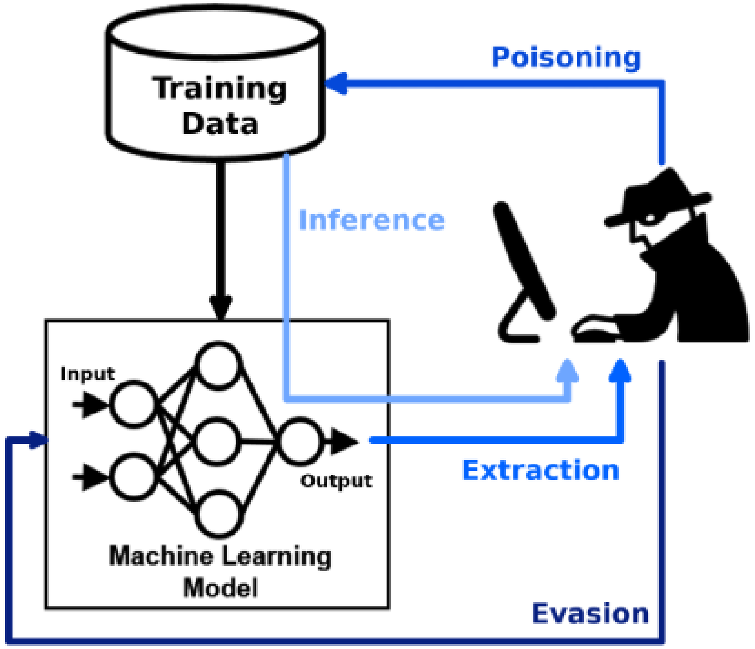
Experiment with Adversarial Techniques

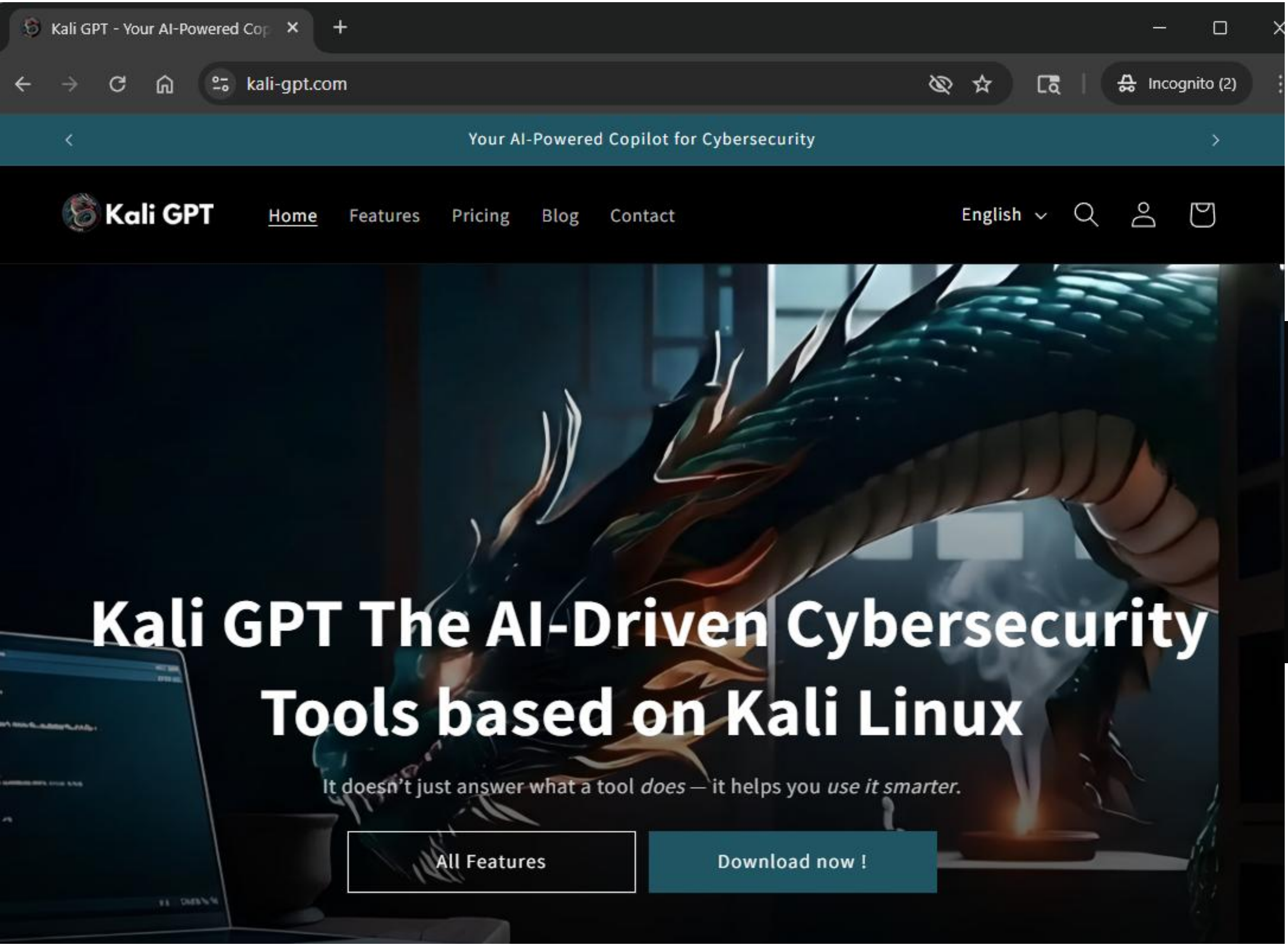
Adversarial Robustness Toolbox

Adversarial Robustness Toolbox (ART) is a Python library for Machine Learning Security. ART provides tools that enable developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all popular machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), all data types (images, tables, audio, video, etc.) and machine learning tasks (classification, object detection, generation, certification, etc.).

The code of ART is on GitHub and the Wiki contains overviews of implemented attacks, defences and metrics.

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

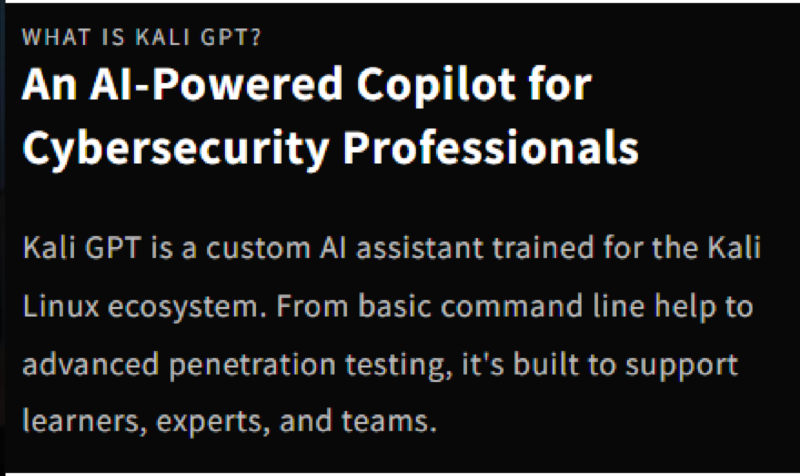




Your AI-Powered Copilot for Cybersecurity

Kali GPT The AI-Driven Cybersecurity Tools based on Kali Linux

It doesn't just answer what a tool *does* — it helps you *use it smarter.*



WHAT IS KALI GPT?

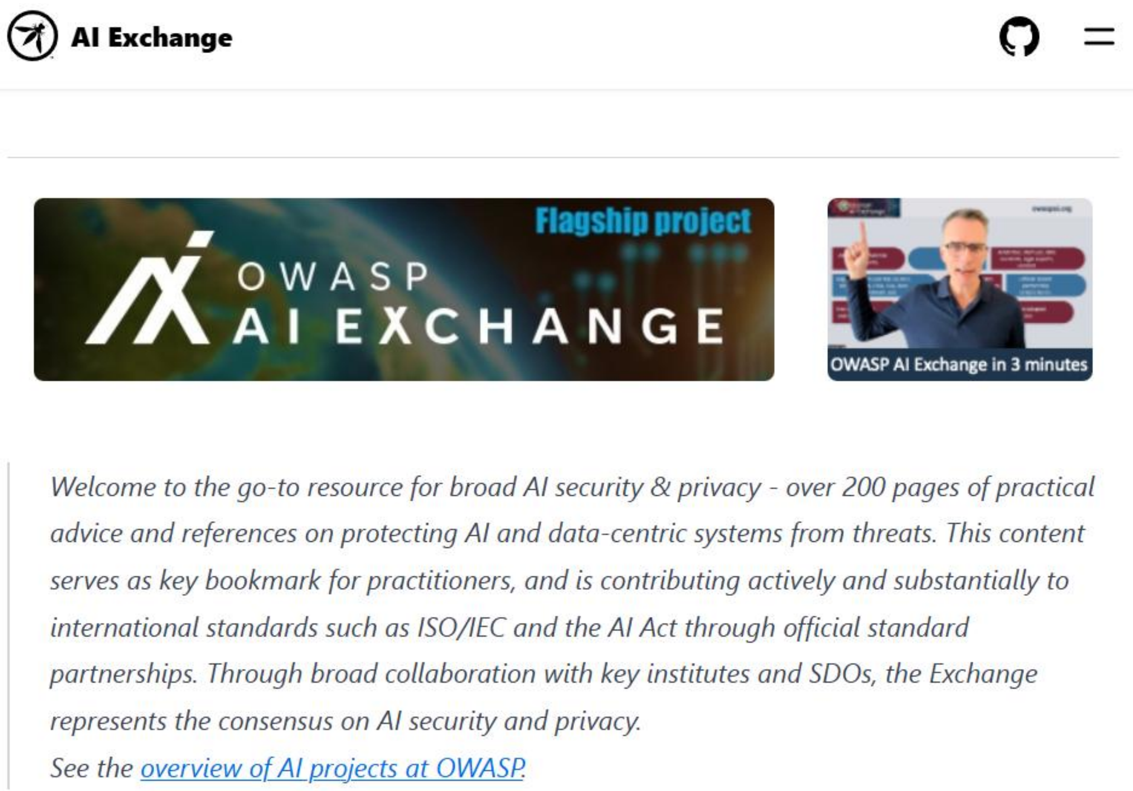
An AI-Powered Copilot for Cybersecurity Professionals

Kali GPT is a custom AI assistant trained for the Kali Linux ecosystem. From basic command line help to advanced penetration testing, it's built to support learners, experts, and teams.

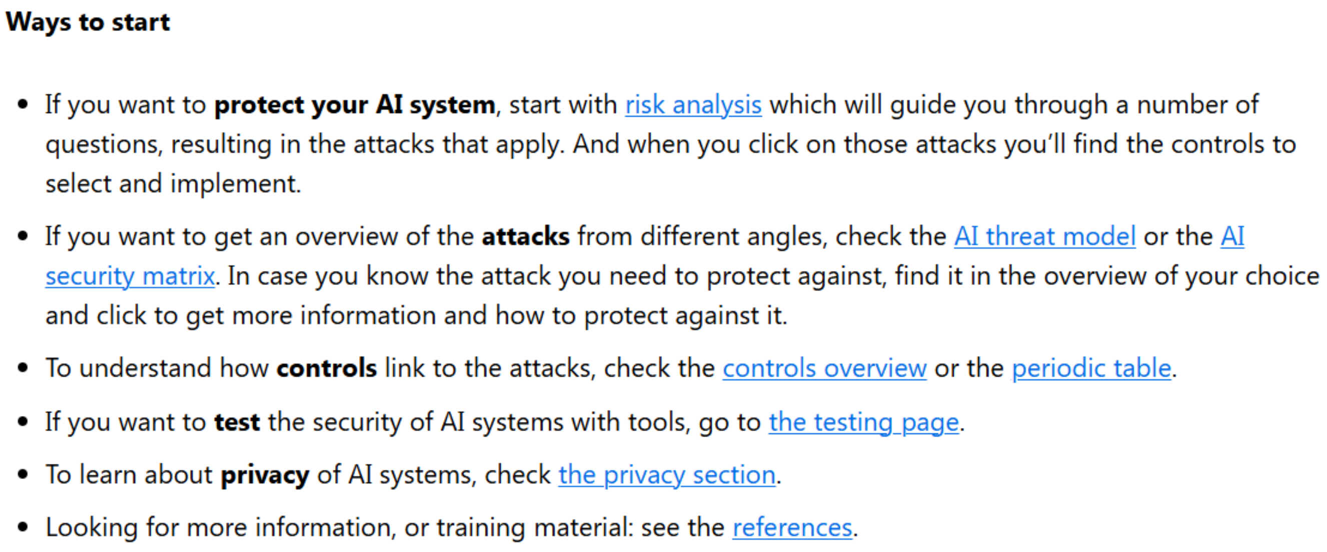
https://kali-gpt.com/

OWASP AI Projects

https://owaspai.org/



> *Welcome to the go-to resource for broad AI security & privacy - over 200 pages of practical advice and references on protecting AI and data-centric systems from threats. This content serves as key bookmark for practitioners, and is contributing actively and substantially to international standards such as ISO/IEC and the AI Act through official standard partnerships. Through broad collaboration with key institutes and SDOs, the Exchange represents the consensus on AI security and privacy.*
>
> *See the overview of AI projects at OWASP.*

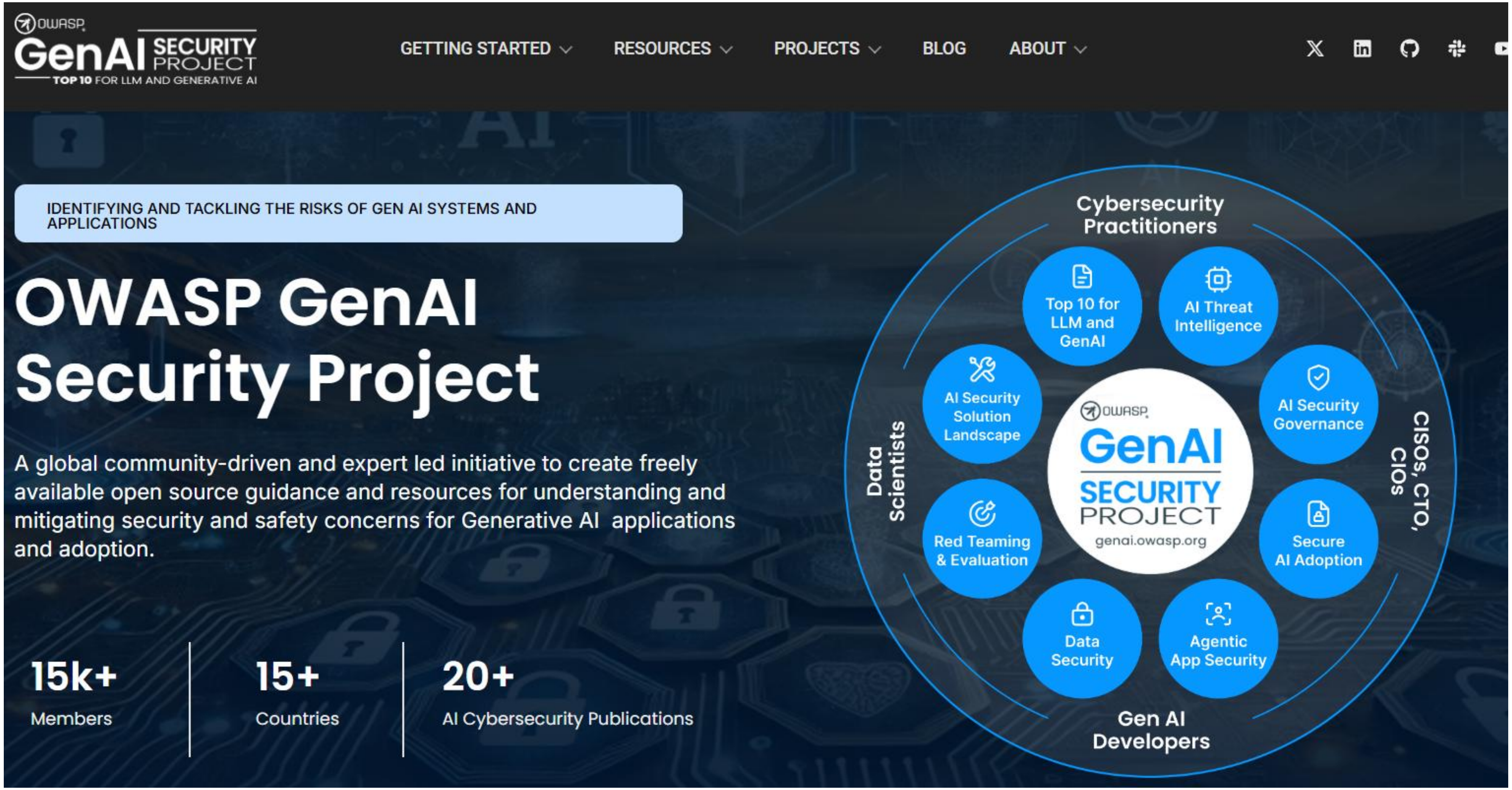


Ways to start

- If you want to **protect your AI system**, start with risk analysis which will guide you through a number of questions, resulting in the attacks that apply. And when you click on those attacks you'll find the controls to select and implement.
- If you want to get an overview of the **attacks** from different angles, check the AI threat model or the AI security matrix. In case you know the attack you need to protect against, find it in the overview of your choice and click to get more information and how to protect against it.
- To understand how **controls** link to the attacks, check the controls overview or the periodic table.
- If you want to **test** the security of AI systems with tools, go to the testing page.
- To learn about **privacy** of AI systems, check the privacy section.
- Looking for more information, or training material: see the references.

OWASP AI Projects

https://genai.owasp.org/



IDENTIFYING AND TACKLING THE RISKS OF GEN AI SYSTEMS AND APPLICATIONS

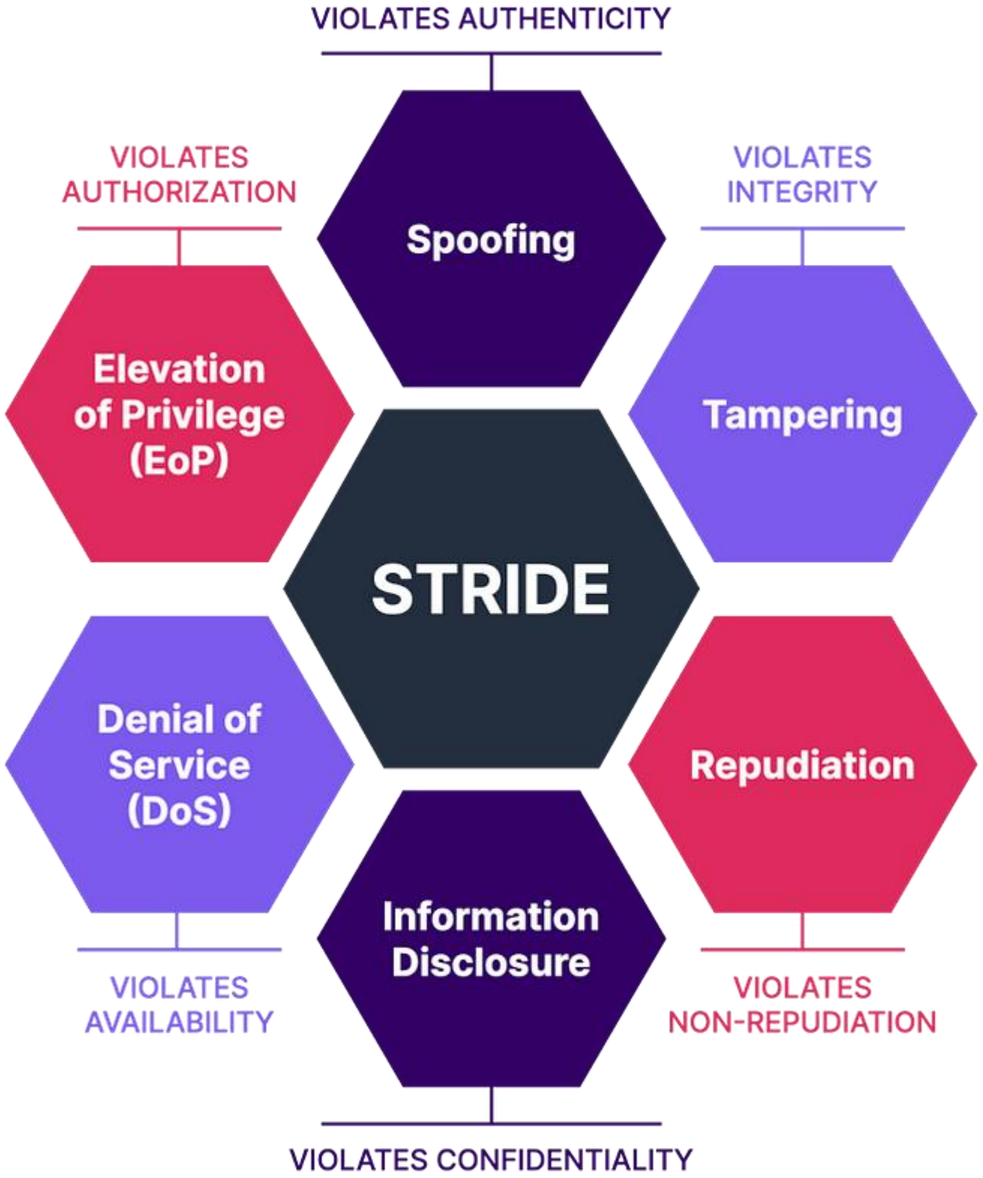
OWASP GenAI Security Project

A global community-driven and expert led initiative to create freely available open source guidance and resources for understanding and mitigating security and safety concerns for Generative AI applications and adoption.

15k+	15+	20+
Members	Countries	AI Cybersecurity Publications

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?



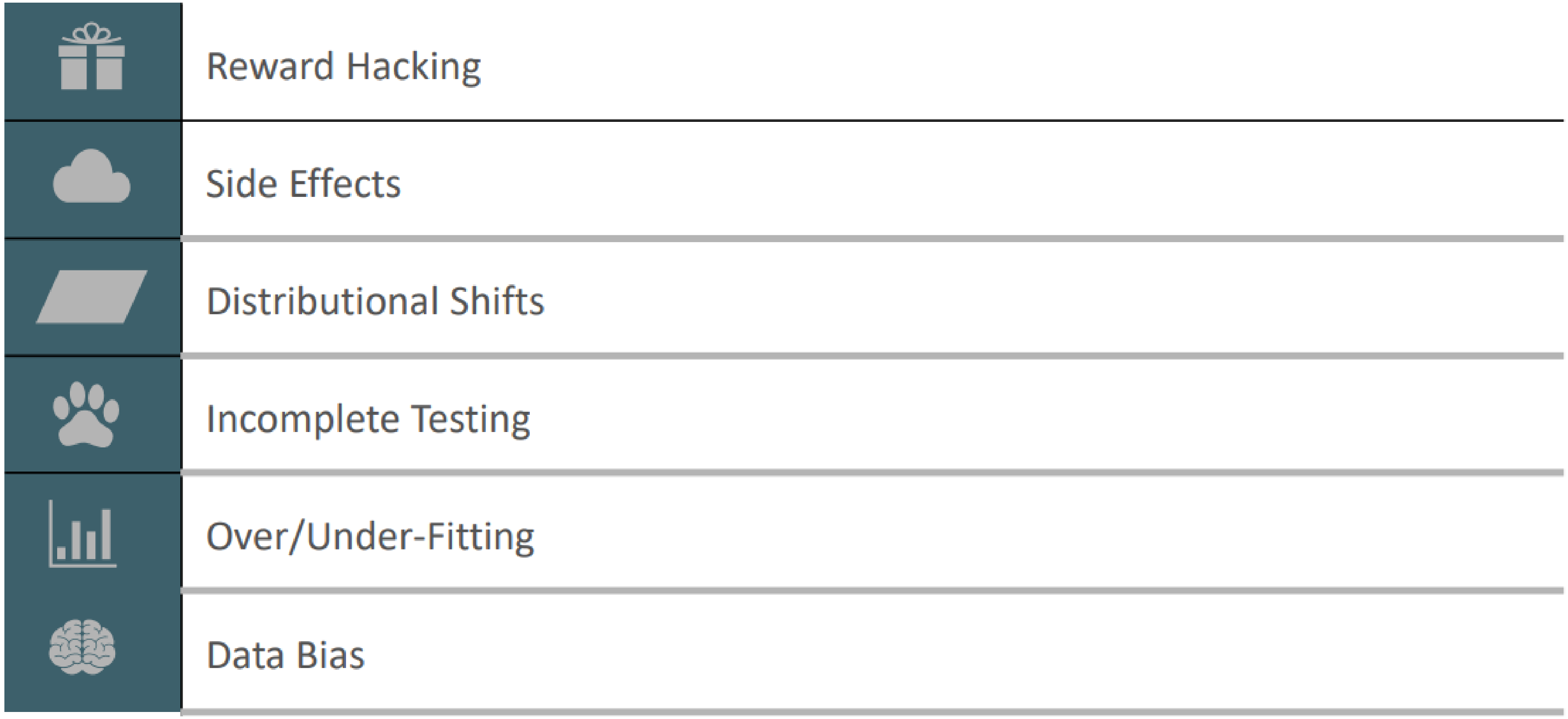
[1] **Shostack, A.** (2014). *Threat modeling: Designing for security*. http://ci.nii.ac.jp/ncid/BB16065709

[2] https://shostack.org/resources/threat-modeling

[3] https://shostack.org/blog/category/threat-modeling

[4] https://www.threatmodelingmanifesto.org/

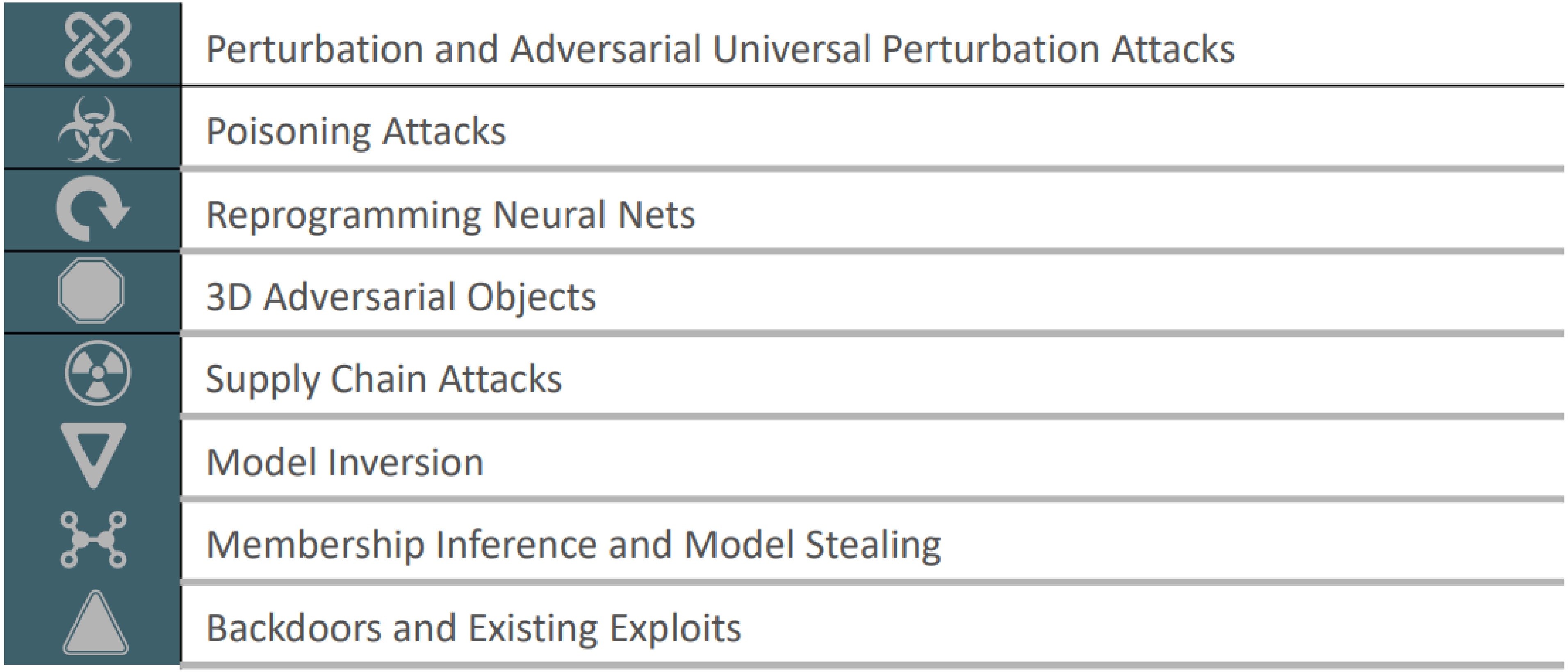
ML Failure Modes – Unintentional Failure



- 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

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

Zero Trust Architecture

- Assume compromise
- Continual verification
- Least privilege

The path to zero trust is an incremental process that may take years to implement.

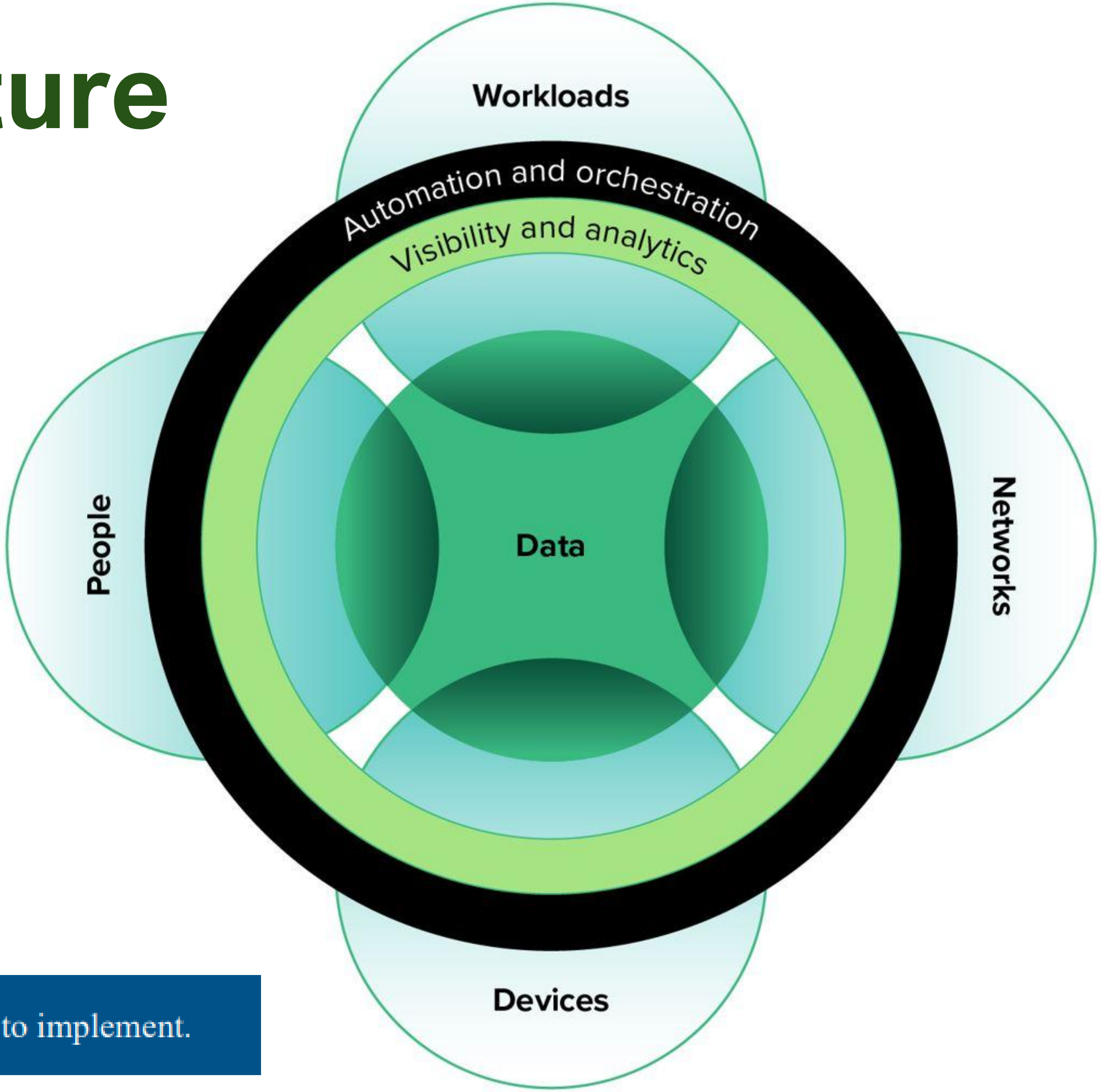


Image source: Zero Trust Security: The Business Benefits And Advantages (forrester.com)

FORRESTER®

Next Steps / “Apply” Slide – Summary

Immediate (the next week):

- Follow Zen-inspired thoughtfulness, fundamentals, & balance
- Review and try 2-3 resources from this presentation
- Set a calendar reminder for 1 month…

Long-term (the next 1-2 months):

- Review the deck and Experiment with 2-3 new resources
- Share with others
- Lead with curiosity, defend with wisdom

Thank You!

For More Information:

ronw@cyber-aaa.com

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

https://linktr.ee/cyberron

MANIFEST YOUR INNER CYBER SUPERHERO!

Bonus Slides

If we have time or when someone asks a question

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***: AI-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.

Cultivating AI Security Awareness Culture

🔐 1. Emphasize Human-Centric Risk Awareness

AI threats often exploit human vulnerabilities — phishing, social engineering, and deepfake manipulation are increasingly AI-powered. Awareness programs must go beyond compliance and focus on behavioral change.

Why it matters: 82% of breaches involve the human element.

🧠 2. Upskill Teams in AI Competencies

AI threats often exploit human vulnerabilities — phishing, social engineering, and deepfake manipulation are increasingly AI-powered. Awareness programs must go beyond compliance and focus on behavioral change.

Why it matters: 82% of breaches involve the human element.

🛡️ 3. Implement AI Governance and Ethical Oversight

AI systems can introduce bias, amplify vulnerabilities, or be misused. Establish clear policies for responsible AI use, including transparency, accountability, and ethical safeguards.

Why it matters: AI is already being used across organizations — often without formal oversight.

🧩 4. Foster Cross-Functional Collaboration

AI risk awareness isn't just for technical teams. Legal, HR, and leadership must be involved in understanding AI's implications for data privacy, compliance, and workforce impact.

Why it matters: AI affects every layer of the organization, from hiring to data governance.