

DOCENT | studios

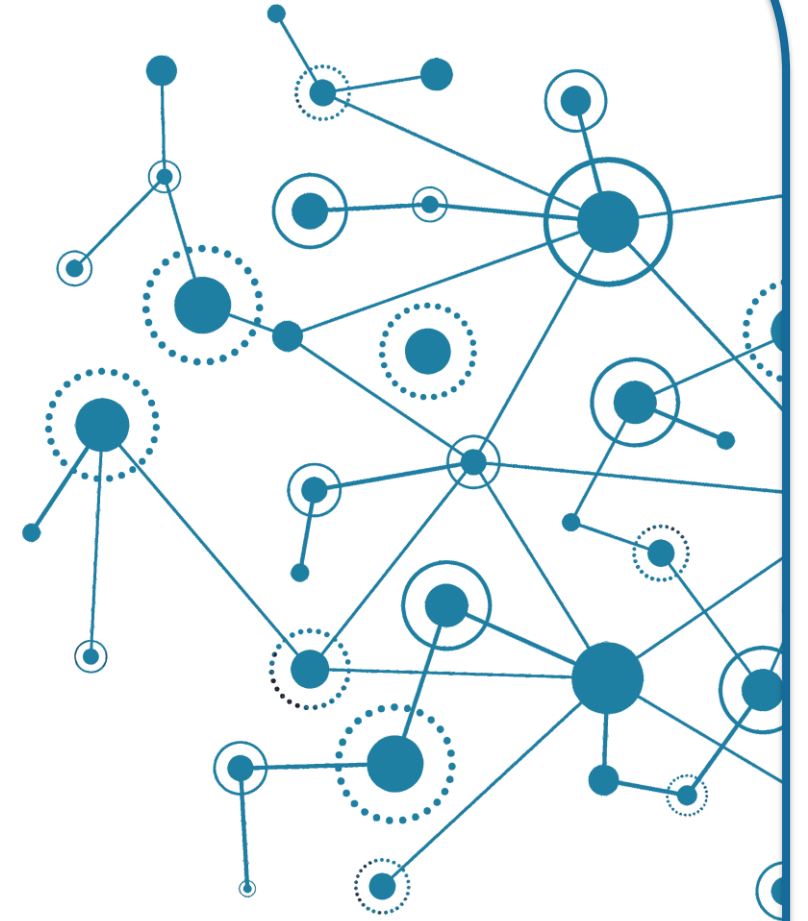


Zen and the Art (and Science) of AI Security

Ron Woerner, CISSP, CISM

vCISO, Consultant, and Educator
Cyber-AAA, LLC

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



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!



Whoami

Ron Woerner

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

LinkedIn:

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

Ron Woerner



TEDx Omaha, 2019,
[Hackers Wanted](#)

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



2025 CORNCON
CYBERSECURITY
CONFERENCE

ZEN AND THE ART OF SECURING AI

Exploring the philosophical and practical dimensions of AI in cybersecurity



AI generated image

Learning objectives:

1. **Explore the dual nature of AI in cybersecurity**, inspired by Zen principles – balancing innovation with introspection to understand both its risks and rewards.
2. **Identify and mitigate malicious uses of AI**, including social engineering, deepfakes, and adversarial attacks, through threat modeling and secure development practices.
3. **Apply philosophical and technical insights** from cybersecurity history to foster resilience, ethical awareness, and proactive defense in AI-driven environments

Exploring the philosophical and practical dimensions of AI in cybersecurity.

The 7 Habits of Highly Effective People – Stephen Covey

The Maturity Continuum®

Each habit is based on universal principles and paradigms of effectiveness, with practices that move learners from dependence and independence to interdependence.

Private Victory® | Habits 1–3

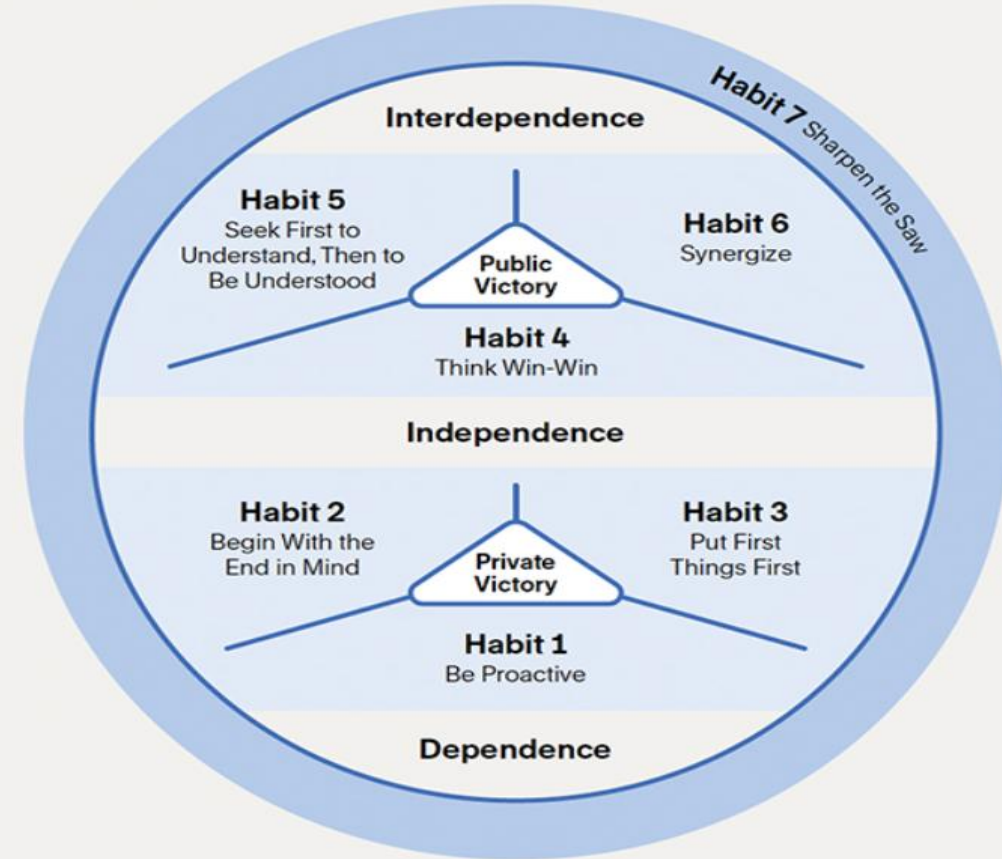
Emphasizes personal mastery, defining desired outcomes and focusing efforts to lay the internal groundwork for success.

Public Victory® | Habits 4–6

Fosters collaboration and synergy with others, building strong relationships that allow us to accomplish more together than we could alone.

Renewal | Habit 7

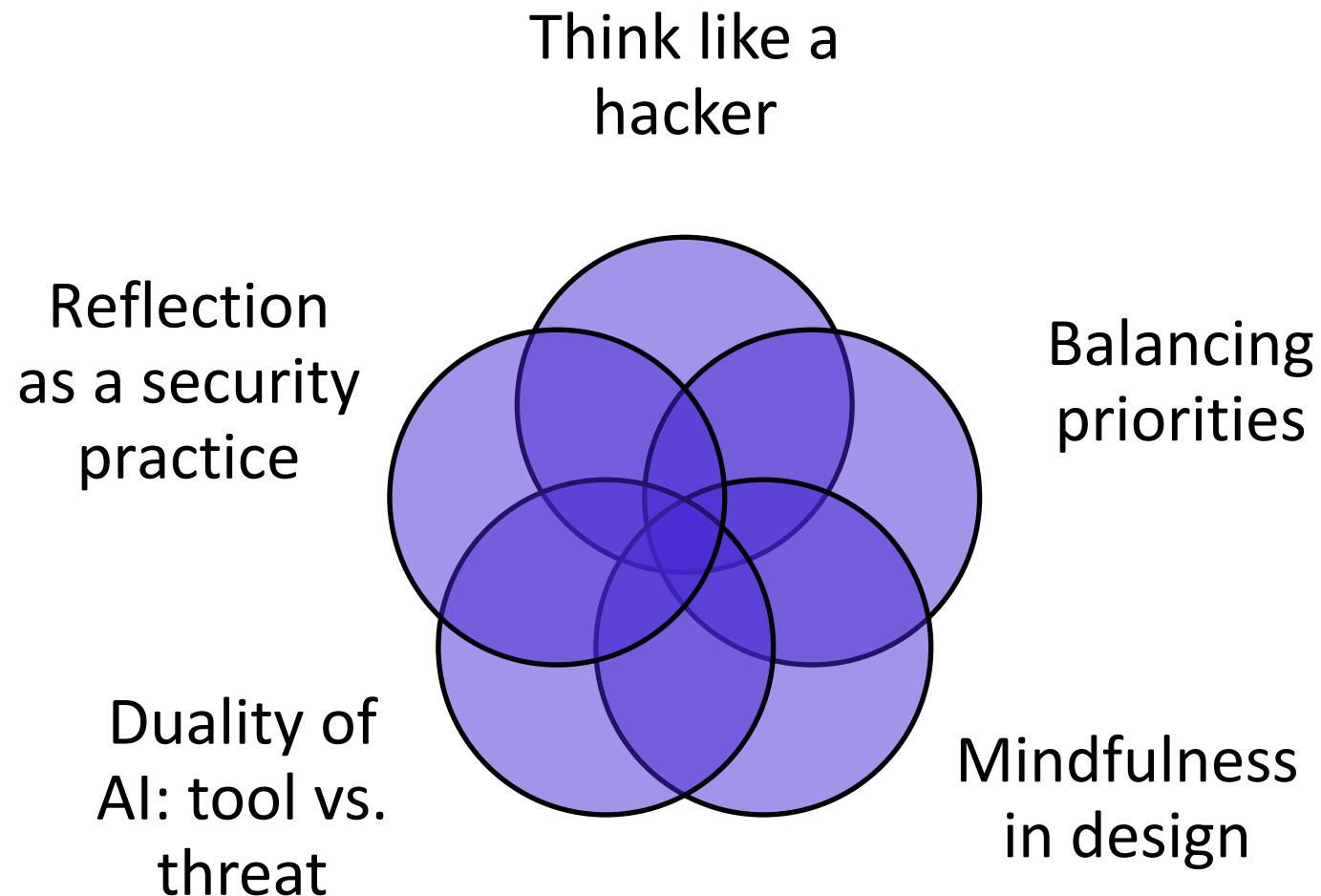
Emphasizes the need for self-renewal and continuous improvement to succeed in and sustain the other habits.



Source: <https://www.franklincovey.com/courses/the-7-habits/>

Zen Principals in AI & Cybersecurity

Finding peace in volatility

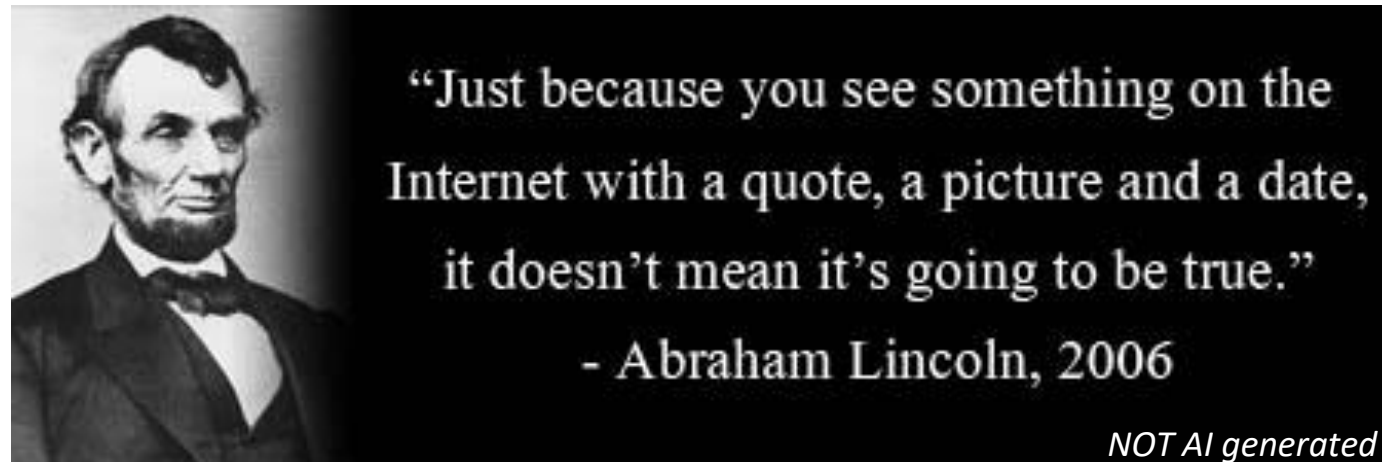


Zen Principals in AI & Cybersecurity

Trust, AND Verify

“Never trust a single data point.”

from my life as a military intelligence officer



Philosophical Reflection

- How do you know what's “true?”
- 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
- Embrace continuous learning and interdisciplinary fluency

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

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.

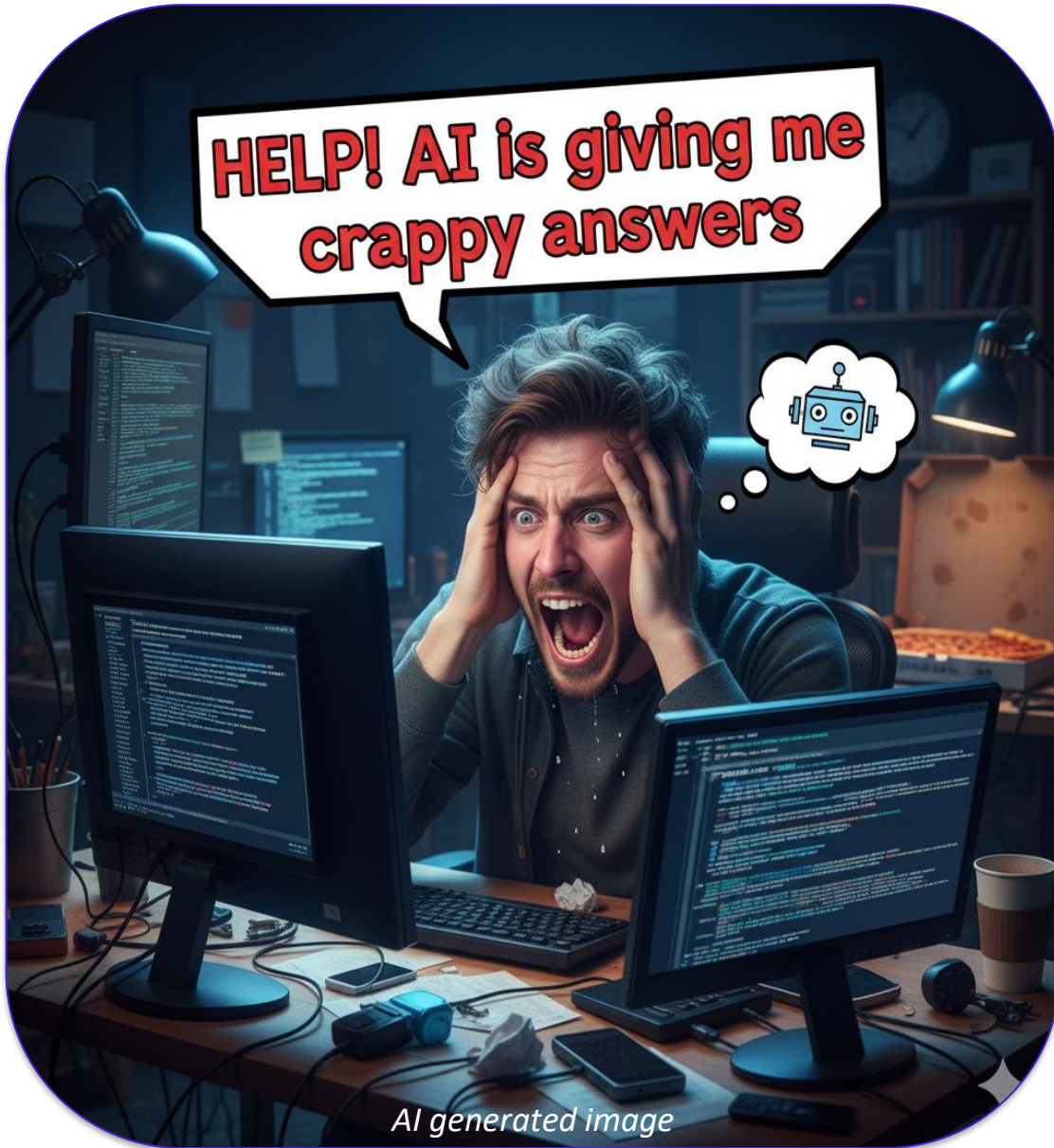


Constraints: What to avoid or include (e.g., "Keep it under 100 words," "Do not include personal names").



Iterate & Refine: AI is a dialogue; refine prompts for better results.

What are your prompting tips?



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

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."



AI 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

Top Questions about AI in Cybersecurity FORRESTER®

- **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/>

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&	Exfiltration
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 techniques
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 LLM 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



ChatGPT



Gemini



perplexity



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




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:


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


Richard P. Feynman

Hatless1der – Ultimate OSINT Collection

**Hatless1der**▼

The Ultimate OSINT Collection


startme

By Hatless1der

🔍 Hatless1der

WELCOME!

This page is for anyone trying to find their way in the overwhelming world of open-source intelligence. It's a collection of my favorite OSINT resources, and I hope it helps you find new ways to learn from some amazing people.



★ Anywhere you see a star, that indicates it's one of my favorites!

If you're new to the amazing world of OSINT, you might start by watching this free video by Micah Hoffman and I at My OSINT Training:
<https://www.myosint.training/courses/introduction-to-osint>

I love networking with the OSINT community, learning with the many amazing people out there, and in recent years have focused heavily on using my experience to make a difference in the world and help others find new ways to learn and grow.

- > OSINT Training: <https://myosint.training>
- > Blog: <https://hatless1der.com>











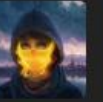



















🌐 [linkedin.com/in/griffin-g](https://www.linkedin.com/in/griffin-g)

🐦 @hatless1der










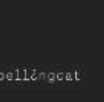




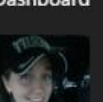



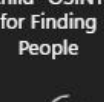

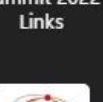

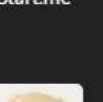
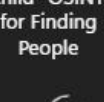






📄 ncptf.org

🔗 clicksafeintelligence.com

News & Blogs

 ★ CQcore - Ginger_T	 ★ cyb_det... Thread Reader App	 ★ hatless1... Blog	 ★ NixIntel	 ★ Sector035 & Week In OSINT	 Bellingcat
 Benjamin Strick Blog	 BushidoTok...	 Dutch OSINT Guy Nico	 Exposing the Invisible	 Gralhix OSINT & GEOINT Tutorials	 Hakin9 Blog
 IntelTechni... Blog	 Key Findings Blog - MW OSINT	 Krebs on Security	 Maltego - OSINT Blogs	 Offensive OSINT	 OH SHINT Blog
 OSINT Combine Blog	 OSINT TEAM Blog	 OSINTCurio...	 osintme.com	 Petro Cherkasets - Medium	 Renato Trabucco Digital Forensics & OSINT Analysis
 Sejuice - OSINT	 seintpl/osint translated	 Skopenow Blog	 Techjournal... Medium	 The Security Neph	 Trace Labs

Tool & Resource Collections

 ★ My OSINT Training's Tools	 ★ cyb_det... tools collection	 ★ IntelTec... OSINT Online Search Tool	 ★ Technise... Tools	 AaronCTI's Online Resources	 AML - Tra
 AsINT_Coll...	 Aware Online OSINT tools	 BBC Africa Eye - Forensics Dashboard	 Bellingcat's Online Investigation Toolkit	 CSE Utopia - Google Custom Searches	 CTI - Paran
 CyberSecStu: Set it on child - OSINT for Finding People	 Dating apps and hook-up sites - frenchPI	 DFIRDetecti... SANS OSINT Summit 2022 Links	 Digintel - CNTY USA	 Digintel OSINT Start.me	 Di
 Exploit Database - Google Hacking	 FBI Tools by Daniel Durnea	 GIUN - Online Research Tools	 Google Advanced Search Operators	 Hun-OSINT Start.me	 i-inte har (2)
 IntelGist's	 Intelligence X	 P4 Machievelli	 juui/aweso	 Journalist's	 Lora

<https://start.me/p/DPYPMz/the-ultimate-osint-collection>

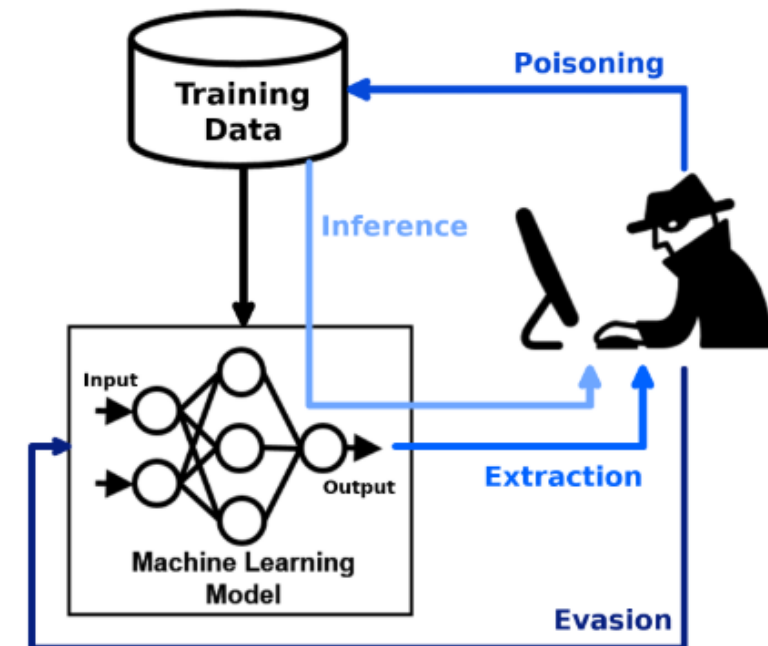
Experiment with Adversarial Techniques*









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](https://github.com/Trusted-AI/Adversarial-Robustness-Toolbox) and the Wiki contains overviews of implemented [attacks](#), [defences](#) and [metrics](#).

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











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

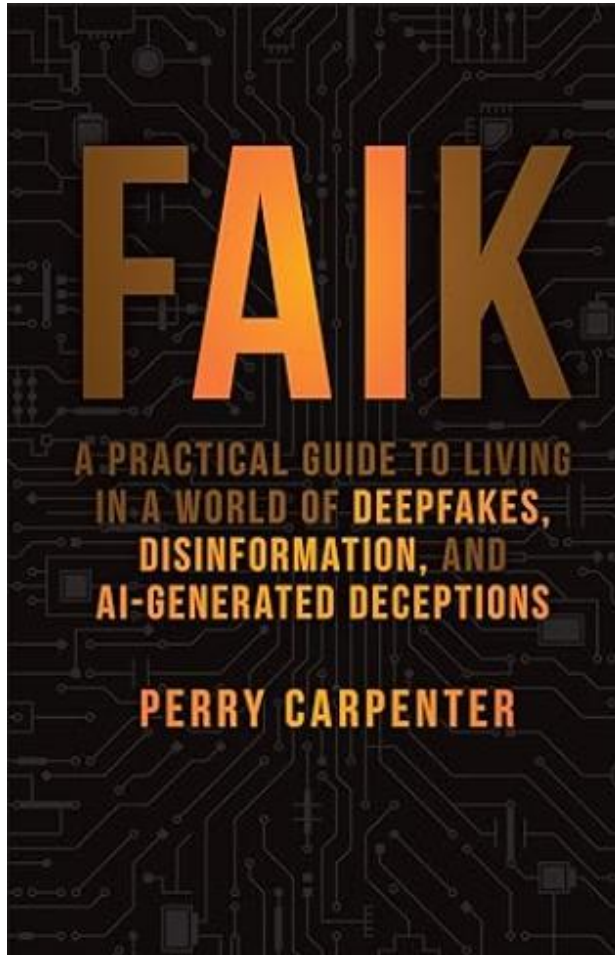
Kumar, et.al. (2022, November 2). *Failure modes in machine learning*. Microsoft Learn.
<https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning>

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



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

Ron Woerner



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, and ...more

thisbookisfaik.com and 4 more links

Subscribe

Home Videos Shorts Podcasts Playlists Posts

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-Powered #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

A-B-C = Always Be Curious

Secure by Design /
Build Security In

Threat Modeling

DevSecOps

Zero Trust



2025 CORNCON
CYBERSECURITY
CONFERENCE

A-B-C = Always Be Curious



2025 CORNCON
CYBERSECURITY
CONFERENCE

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

Trustworthy & Responsible AI Resource Center

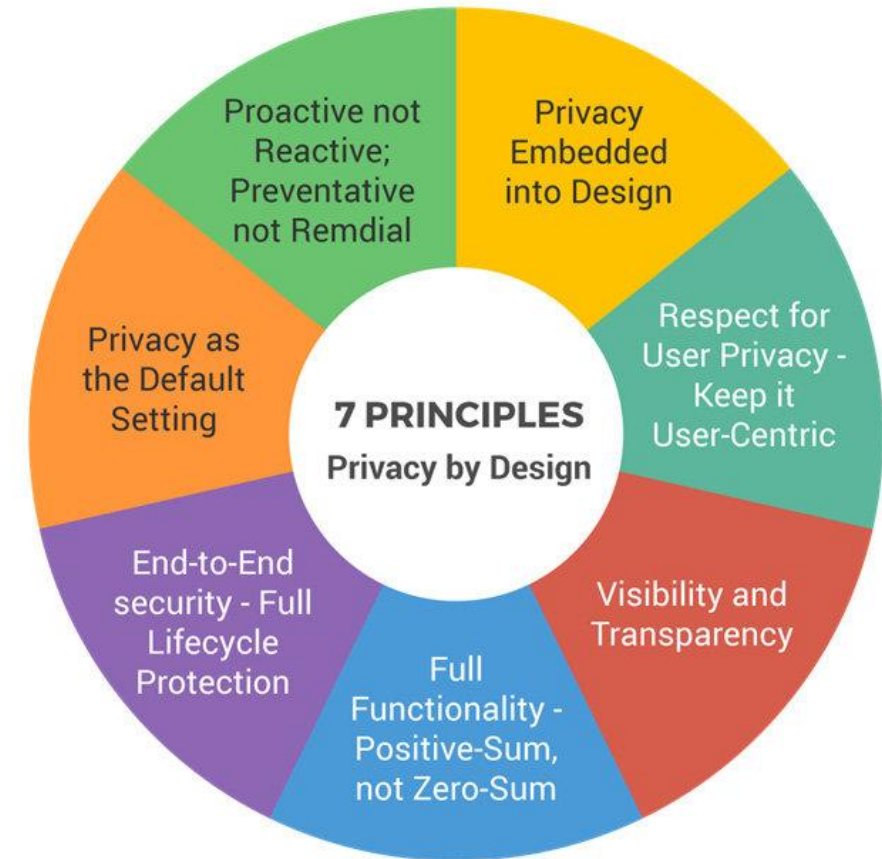
NIST AI Risk Management Framework

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>

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



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>

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>



<https://www.cisa.gov/securebydesign>

OWASP AI Projects

<https://owaspai.org/>



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](#).



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](#).

OWASP AI Projects

<https://genai.owasp.org/>

OWASP GenAI Security Project
TOP 10 FOR LLM AND GENERATIVE AI

GETTING STARTED ▾ RESOURCES ▾ PROJECTS ▾ BLOG ABOUT ▾

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+
Members

15+
Countries

20+
AI Cybersecurity Publications

GenAI SECURITY PROJECT
genai.owasp.org

Cybersecurity Practitioners

- Top 10 for LLM and GenAI
- AI Threat Intelligence
- AI Security Governance
- Secure AI Adoption
- Agentic App Security
- Data Security
- Red Teaming & Evaluation
- AI Security Solution Landscape

Data Scientists

CISOs, CTO, CIOs

Gen AI Developers



Guidelines for secure AI system development

Executive summary.....	5
Introduction.....	6
Why is AI security different?.....	6
Who should read this document?.....	7
Who is responsible for developing secure AI?.....	7
Guidelines for secure AI system development.....	8
1. Secure design.....	9
2. Secure development.....	12
3. Secure deployment.....	14
4. Secure operation and maintenance.....	16
Further reading.....	17



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

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.

Zen Principals in AI & Cybersecurity

Last thought:



AI generated image

“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



2025 CORNCON
CYBERSECURITY
CONFERENCE

2025 CORNCON CYBERSECURITY CONFERENCE



For More Information:

ronw@cyber-aaa.com

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

<https://linktr.ee/cyberron>



MANIFEST YOUR INNER CYBER SUPERHERO!



2025 CORNCON
CYBERSECURITY
CONFERENCE



2025 CORNCON
CYBERSECURITY
CONFERENCE

2025 CORNCON CYBERSECURITY CONFERENCE

MANIFEST YOUR INNER CYBER SUPERHERO!



“Apply” Slide

- Bullet point here (see slides 7 – 10 for instructions)
- Bullet point here
- Bullet point here



2025 CORNCON CYBERSECURITY CONFERENCE

MANIFEST YOUR INNER CYBER SUPERHERO!

