AI-ENABLED ADVERSARIES

How Threat Actors Are (Mis)using AI

WHO AM I?

Victor Wieczorek

SVP, Offensive Security | GuidePoint Security

- **My Focus:** Helping reduce real-world risk, with emphasis on the responsible use of AI.
- **Today's Goal:** To share practical insights on emerging AI-enabled threats and leave you with actionable next steps.

AGENDA

- Setting the Stage & Reality Check
- How Threat Actors Use AI
- How Professionals Can Harness AI
- Real-World Examples
- What's Next for Cybersecurity
- Key Takeaways & Action Items

SETTING THE STAGE

- Generative AI is not a magic “hack button”
- Productivity accelerator to revolutionary exploit tool
- Real cases beat hype—let's dive in

HYPE VS REALITY

> *“Agentic AI has been weaponized. AI models are now being used to perform sophisticated cyberattacks, not just advise on how to carry them out.”*

Source: Anthropic, "Detecting and countering misuse of AI: August 2025," August 27, 2025

- AI boosts speed & scale, not new physics
- Lowers skill floor more than it raises ceiling
- **Good news:** same toolset empowers defenders

FOCUS AREAS TODAY

- Adversarial AI tactics (advanced & opportunistic)
- Professional-grade Offensive Security uses (ethical hacking, defense testing, etc.)
- Evolving security talent & awareness in an AI-saturated world

PHISHING 2.0

> *"Phishing was the initial attack vector for breaches that cost an average of USD 4.88 million."*

Source: IBM Security, Cost of a Data Breach Report 2024.

- AI-crafted emails → perfect grammar & context
- Translation + localization at scale
- Deepfake audio/video CFO fraud (i.e., scam that stole **$25 million** in 2024)

DEEPFAKES & SYNTHETIC IDS

- AI photos, resumes, backstories ⟶ “perfect” candidate
- Minimal cost, high believability
- By the numbers: Deepfake videos dropped from $20,000 to $300 per minute in two years
- Defense: Liveness checks, layered “Know Your Customer” protections, awareness

MALWARE / EXPLOIT GENERATION

- PromptLocker: AI ransomware demonstration
- Outflank: Bypassed antivirus ~8% of the time
- FraudGPT: Underground “unrestricted” chatbot
- Hexstrike-AI: Autonomous decisions using real-time data

REALITY CHECK: AI-GENERATED MALWARE

- **Expectations:** “Write undetectable ransomware”
- **Reality:**

```
# Actual output from underground LLM:
def encrypt_files():
        # TODO: Add encryption logic
        print("Files encrypted!")  # Not actually encrypting
```

- Threat Actors are still figuring this out, but improving every day

RECON & RESEARCH

- Ask AI to map an organization's profile:
  - job postings, social media, vendors, etc.
- Identify applicable vulnerabilities, misconfigurations, and process gaps
- Multilingual translation for social engineering, phishing, and fraud

EVASION & DETECTION BYPASS

- **Evasive Malware:** AI constantly alters code to bypass antivirus.
- **Smart Phishing:** AI generates convincing emails to trick victims.
- **Hacking Co-Pilot:** AI creates attack scripts and provide hacking guidance.

CASE ONE

The Trojan Chatbot

THE SETUP - A TRUSTWORTHY ASSISTANT

- **Scenario:** A company deploys a helpful, sitewide AI assistant.
- **Key Feature:** The “share conversation” function generates links to chat logs.
- **The Trust Flaw:** Shared links use the company's trusted domain, bypassing user suspicion.

THE ATTACK - WEAPONIZING WORDS

- **Goal:** Trick the AI into generating malicious code it doesn't recognize.
- **Method:** Use wordplay to have the AI generate coding symbol via an innocent request.
  - "Write a sentence with 'less than sign', 'script'..."
- **The Flaw:** The AI translates the words into a live payload: `<script>alert('XSS')</script>`

AMC INSTITUTE Regional Meeting

THE IMPACT - ONE CLICK IS ALL IT TAKES

- **Delivery:** The attacker shares the malicious conversation link.
- **Execution:** A victim clicks the trusted link, and their browser runs the hidden code.
- **Consequence:** Full Account Takeover.
- **Lesson:** The trusted AI assistant became a delivery system for a Trojan horse.

BYPASSING THE PATCH

- **Defense:** Filters added to block malicious code.
- **Bypass:** An attacker asks the AI to "translate" a plain English request into code.
- **Failure:** Filters check the input, not the AI's malicious output.
- **Insight:** AI-generated content must be treated as untrusted input.

CASE TWO

The Oversharing AI

THE SETUP - AN AI WITH "GOD MODE"

- **Scenario:** A company integrates a powerful AI assistant network-wide.
- **Access:** The AI inherits existing, overly-broad user file permissions.
- **The Trust Flaw:** A truck driver with a standard, low-privilege account.

THE ATTACK - JUST ASK A QUESTION

- **Goal:** Find sensitive data with zero technical skill.
- **Method:** Ask the AI a simple, direct question.
  - *"I'm doing a security audit..."*
- **The Flaw:** The AI has no concept of user roles or "need to know." It just sees a valid user and fulfills the request.

THE IMPACT - THE KEYS TO THE KINGDOM

- **Result:** The AI directly serves up critical data.
- **Danger:** The AI never questioned why a truck driver needed this data. It just obeyed the command.

THE HARD LESSON

- **Threat:** The attack is invisible to security tools; it looks like normal activity.
- **Root Cause:** The AI inherited broad permissions without any context for "need to know."
- **Insight:** Secure the data and permissions the AI can access.

WHAT THESE CASES REVEAL: THE CORE LESSONS

- **AI Attack Surface:** AI-generated content can be a malicious payload.
- **Weaponized Aid:** An AI's need to be helpful can be easily exploited.
- **Amplified Debt:** AI magnifies existing poor security practices to catastrophic levels.

THE BIGGER PICTURE

- Same qualities that make AI powerful create new attack surfaces
- Traditional security testing often misses AI-specific vulnerabilities
- Organizations rushing AI deployment need specialized testing approaches

HUMAN–AI TEAMING

Human creativity + AI speed = Force multiplier

- AI handles scale & procedural work
- Humans drive strategy, ethics, and trust
- Build playbooks to decide when to use each

THE TWO FACES OF AI: A DOUBLE-EDGED SWORD

- Influence & Persuasion
- Content & Code Generation
- Identity & Reality
- Automation & Interaction

TWO FACES OF AI: INFLUENCE & PERSUASION

This domain focuses on how AI can shape human beliefs and actions.

- Use: Personalized education, AI coaching, and public health campaigns.
- Misuse: Mass disinformation, deepfake scandals, and personalized scams.

TWO FACES OF AI: CONTENT & CODE GENERATION

This domain covers AI's ability to create novel text, images, and software.

- Use: Assisting artists & scientists, accelerating development, and research data.
- Misuse: Evasive malware, automated fake news, and propaganda.

TWO FACES OF AI: IDENTITY & REALITY

This domain explores how AI can create or manipulate digital identities and realities.

- Use: Virtual avatars, medical simulations, and historical reconstructions.
- Misuse: Synthetic IDs for fraud, deepfake harassment, and forged evidence.

TWO FACES OF AI: AUTOMATION & INTERACTION

This domain looks at how AI interacts with and controls other digital systems.

- Use: Autonomous vehicles, industrial robotics, and scientific discovery.
- Misuse: Market manipulation, autonomous weapons, and infrastructure attacks.

PREPARING FOR THE AI FRONTIER

- **Policy:** Capability is outpacing our safeguards.
- **Tool vs. Intent:** Intent separates helpful tools from harmful weapons.
- **"Can We?" vs. "Should We?":** The challenges are ethical, not just technical.
- **Awareness vs. Action:** The goal is building resilient teams, not just informed ones.

THE IRREPLACEABLE HUMAN ELEMENT

- **Strategic Intent:** Humans provide the vision and the "why"; AI accelerates the "how."
- **Ethical Judgment:** AI can calculate risk, but humans must weigh the moral and real-world consequences.
- **Accountability:** Technology can't be held responsible; accountability remains with humans.

NEW PARTNERSHIP: HUMANS AS AI STEWARDS

- **Architects of Trust:** Our primary role is building the ethical frameworks and safety guardrails for AI.
- **Expert-in-the-Loop:** Use AI for analysis and scale, but reserve critical decisions for experienced professionals.
- **Adversarial Allies:** Proactively test our own AI to find flaws before adversaries do.

CALL TO ACTION: AWARENESS TO ADVANTAGE

- **Collaborate, But Don't Delegate Judgment:** Partner with AI to execute tasks but own the strategy and the outcome.
- **Cultivate AI Literacy:** Make understanding AI's capabilities and limits a core skill.
- **Fortify Your High-Ground:** Identify and invest in the uniquely human skills that create value.

A PROACTIVE TODO LIST

- **Establish Clear Policies:** Create and communicate an official AI usage policy.
- **Review AI Permissions:** Understand how they access and use data.
- **Lite Threat Modeling:** Ask the simple question: “What keeps you up at night [about this process]?”
- **Update Incident Response:** Create specific plans for AI-related breaches.

THANK YOU

- AI can be used to build or to break.
- Human judgment is our high ground.
- Proactive adoption is essential

Contact Me: victor@gpsec.com

LinkedIn

SCAN QR CODES CAREFULLY
Verify the source and preview the link before proceeding