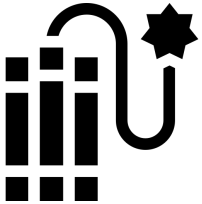


Dual Use: Cyber Offense vs. Defense

Accountability in AI (04 Jun 2026)

Yannis Hofmann

Definition of Dual Use

Definition

Dual-use items refer to equipment, machines, goods and technology that can be used for both civilian and military applications.

Examples

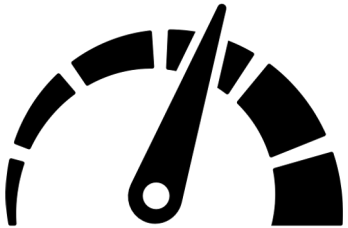
A **Dynamite:** used for mining and war.

B **Nuclear Research:** “clean” energy and war.

C **Image Object Recognition:** used for automation and war.

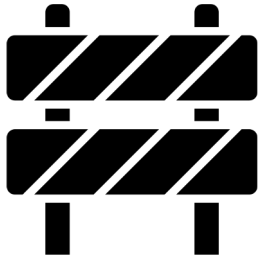
The concept of dual use hard- and software has been around for quite some time.

The same attributes make AI great for offense and defense



Scale & Speed

Autonomy



Lower Skill Barrier

AI can help make systems and companies more secure

Google AI "Big Sleep" Stops Exploitation of Critical SQLite Vulnerability Before Hackers Act

Ravie Lakshmanan Jul 16, 2025

https://thehackernews.com/2025/07/google-ai-big-sleep-stops-exploitation.html

AI can also be used to attack companies on a new scale

https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf

AI can also be used to attack companies on a new scale

ANTHROP\C

Disrupting
first reporte
AI-orchestr
cyber espio
campaign

How AI is Silently Rewriting the Rules of Cyber Attacks

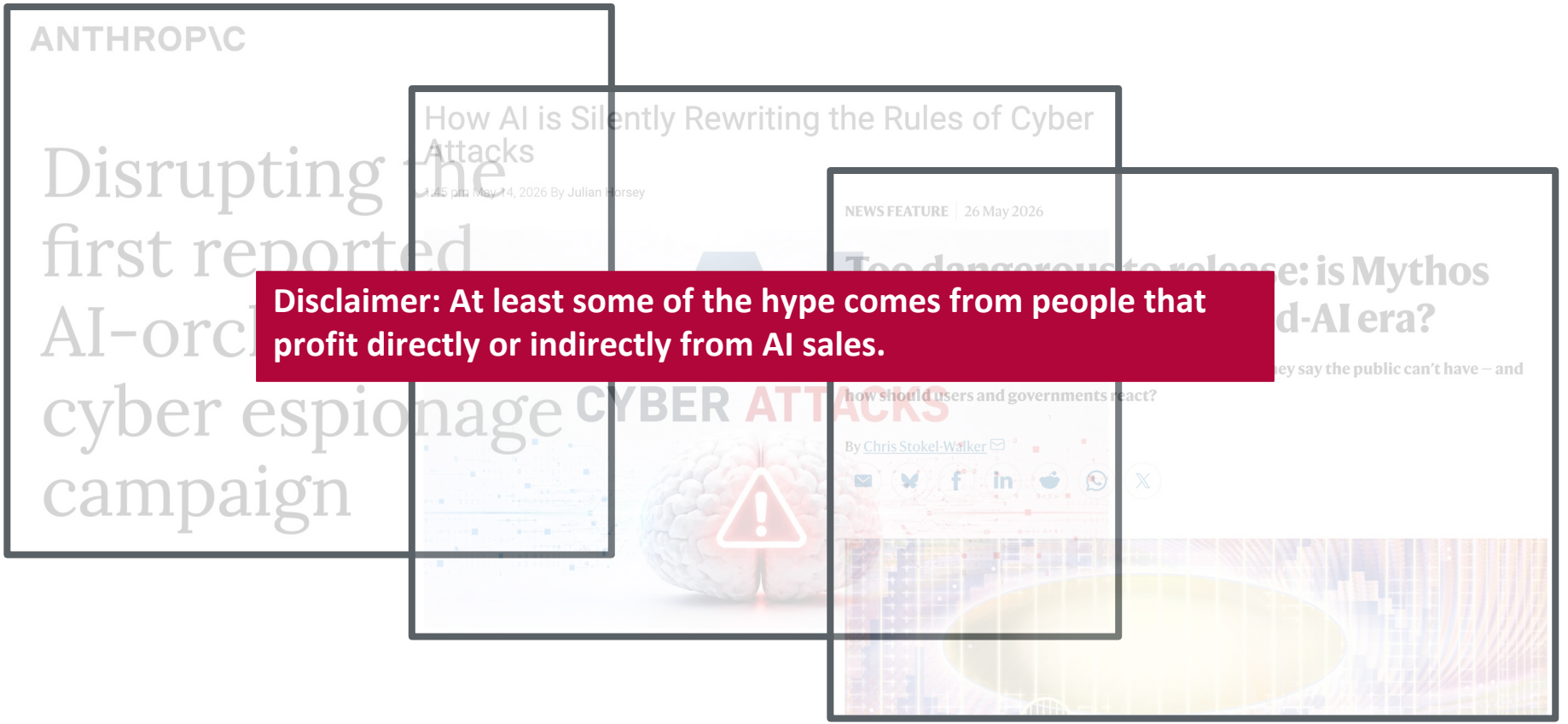
1:45 pm May 14, 2026 By **Julian Horsey**

https://www.geeky-gadgets.com/autonomous-malware-threats-2026/

AI can also be used to attack companies on a new scale

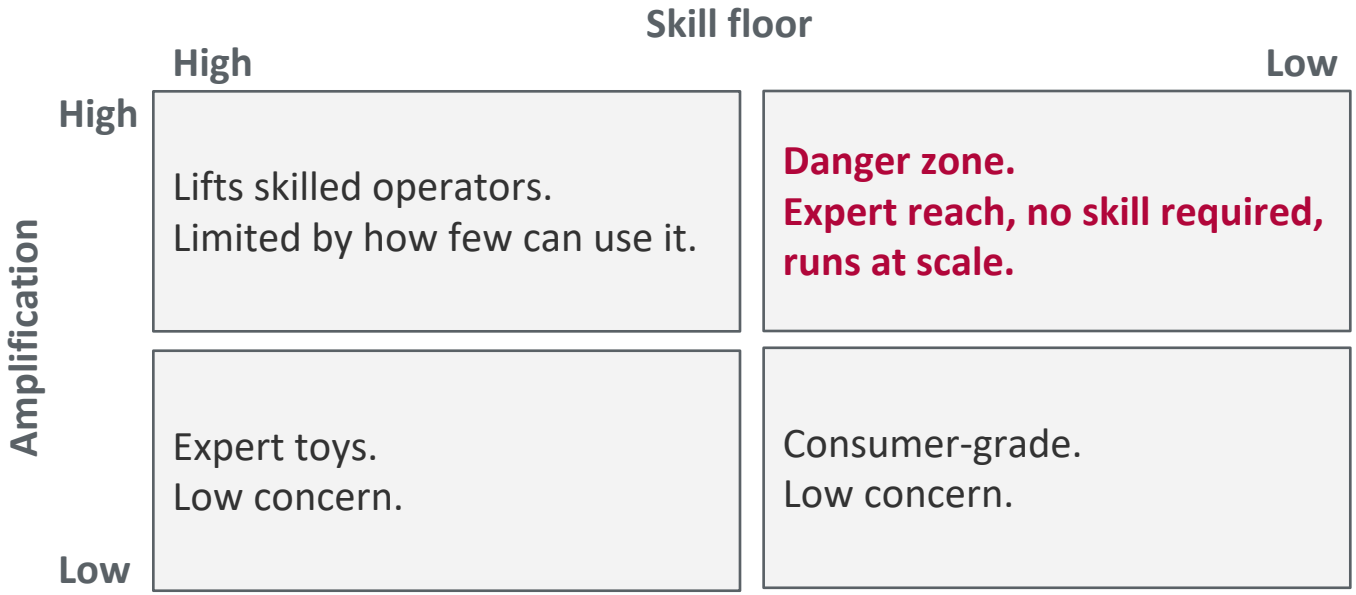
https://www.nature.com/articles/d41586-026-01617-2

AI can also be used to attack companies on a new scale



Disclaimer: At least some of the hype comes from people that profit directly or indirectly from AI sales.

When is AI actually dangerous for cyber attacks?

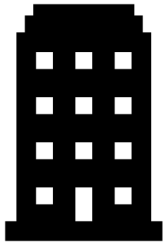


Skill floor — how good you have to be to use it at all

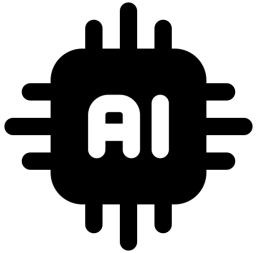
Amplification — once you can use it, how far it takes you

Three options for responsibility

Operator



Provider



AI itself?

Same dual-use, same questions: if Cobalt Strike's makers aren't liable, why would AI's be?

Cobalt Strike®
Adversary Simulations and Red Team Operations

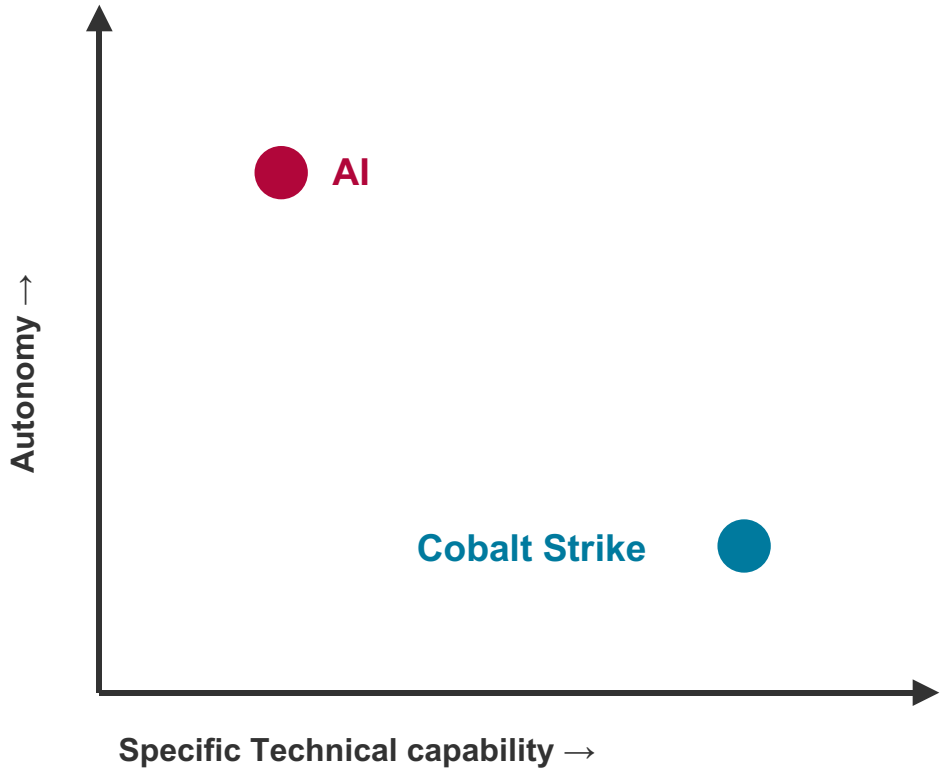
Software for Adversary Simulations and Red Team Operations

Adversary Simulations and Red Team Operations are security assessments that replicate the tactics and techniques of an advanced adversary in a network. While penetration tests focus on unpatched vulnerabilities and misconfigurations, these assessments benefit security operations and incident response.

REQUEST PRICING WATCH A DEMO

https://www.cobaltstrike.com/

The difference: Cobalt Strike brings sharper capability, AI brings autonomy



Paper: Turning LLM weaknesses into cyber defenses

Problem	**Attack volume and quality are rising sharply** as AI-enabled adversaries automate recon, exploitation, and execution
Core idea	**Exploit weaknesses of attacking LLM** such as bias, blind trust in inputs, and limited memory to mislead, stall, or shut it down
Method	**Plant prompt injection and misleads** in the environment to confuse attacking LLMs
Structure	**Taxonomy of 11 strategies** organized under three goals: Prevent, Detect, Delay.

Paper: The Best Defense is a Good Offense: Countering LLM-Powered Cyberattacks Pasquini, Kornaropoulos, Ateniese et al.

Paper: Three defensive goals against AI-powered attacks

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Prevent Stop the agent succeeding	**Detect** Tell an LLM from a human	**Delay** Burn the attacker's time/money
▪ **Convince** it to stop the attack ▪ **Trigger built-in safety** of the upstream model provider ▪ **Lure into reverse shell** back on the attacker	▪ **Plant honeytokens** with invisible characters ▪ **Convince it to execute code** that triggers a detection	▪ **Trap the agent** in endless search loops ▪ **Drain compute** with "sponge" prompts ▪ **Flood it with fake information** like CVEs, credentials, open ports

Paper: Results, benefits and shortcomings

Setup

Black-box tests vs. 4 LLM pentest tools and 4 models

Key results

- **>90% success** for best technique-asset pair
- **Most robust:** Convince to Stop, Overwhelm, Change Objective, Trigger Safeguards
- **Simple bait** works best

Benefits

- **Cheap to deploy**
- **Hard to fix** as it targets fundamental LLM flaws
- **Layering shrinks attacker success drastically** and helps against human attackers too

Shortcomings

- **Agents without reasoning state** mostly resistant against these defenses
- **Tests limited** to mostly single prompt tests
- **Inherently an arms race**
- **Misinformation** can also harm the company itself

Questions for discussion

Risk vs. benefit	How does the added protection and defense capabilities compare to the larger attack surface which often comes with AI deployments?
Liability	Should the company providing the model be held liable as well, or only the operator?
Accountability	When an autonomous AI agent causes harm, how do we trace and audit its decisions to assign responsibility?