THE ALGORITHM

ARMS RACE

AI and the Future of National Security

Privacy + Security Forum | Spring Academy | May 7, 2026 | George Washington University

Eun Young Choi Arnold & Porter | **Cynthia Kaiser** Halcyon.ai | **Sean Newell** DOJ NatSec Cyber

The same AI capabilities built for beneficial purposes have become the defining instrument of both criminal and nation-state cyber threats —

and the legal, regulatory, and defensive frameworks were not built for this moment.

Three-Point Framework for AI

01

The Threat Landscape & How to Defend Against AI Threats

Anticipate and defend against those who use AI to power malicious cyber activity.

02

Defend the Innovators

Protect companies who are integrating AI into their workflows for lawful use – and those who are building next-generation AI – from those who would use these tools to attack systems, or steal the technology to use it against others.

03

Use AI Responsibly

Leverage AI — triaging data, accelerating investigations — with humans in the loop and consistent with law, to strengthen resilience for all.

TOPIC 1

The Threat Landscape — How AI Powers Criminal and Nation-State Cyber Operations

How AI Powers Attacks Across the Full Kill Chain

1 RECON

AI scans massive data sets to identify vulnerabilities, profile targets, scrape OSINT on personnel

2 SOCIAL ENG.

Personalised phishing, deepfake voice/video calls of managers & executives — leverages recon data for precision targeting to gain initial access

3 EXPLOIT DEV

Auto-generates & tailors exploits to specific vulnerabilities identified during recon; use of Mythos-class tools could potentially synthesize working zero-days autonomously

4 LATERAL MOVEMENT AND EXFIL

Navigates networks faster; mimics legitimate user behaviour to evade detection; maps and extracts high-value targets; AI triages exfiltrated data to surface stored passwords and credentials, enabling rapid expansion of access

5 MALWARE EVOLUTION

Self-modifying malware evades YARA rules; AI debugs and improves code autonomously; adapts in real time to defenses

Crime-as-a-Service / AI-as-a-Service: WormGPT, FraudGPT — dark-web LLMs lowering the bar for low-skill criminals. Subscription from hundreds to thousands per month.

AI Used Primarily for Initial Access

Experimenting with AI for Foothold Expansion

Minimal Adoption in Security Bypass, Exfiltration, and Encryption

The Criminal Ecosystem: Key Ransomware Actors & Their AI Adoption

LOCKBIT

THE PRE-AI BENCHMARK

- Most prolific pre-2024 RaaS — 2,000+ victims, $91M in U.S. ransom payments
- Operation Cronos (Feb 2024): FBI/Europol seized infrastructure, decryption keys, 200+ crypto wallets
- Obfuscation (XOR, stack-string, dynamic API resolution) — sophisticated traditional techniques, not AI-driven
- Disruption scattered affiliates into RansomHub, Qilin, others — accelerating AI adoption by successors
- **Lesson: pre-AI ransomware was a franchise model problem — now it is an AI arms race**

RANSOMHUB

AI-FORWARD SUCCESSOR

- Emerged Feb 2024; absorbed LockBit and ALPHV affiliates; 600+ victims in 2024
- 90% affiliate revenue share — highest in RaaS market; attracted elite operators
- Affiliate backdoor shows hallmarks of **AI-assisted code development**: structured classes, robust error handling (Guidepoint Security, Jan 2025)
- RaaS breakout time: 48 min. (2024) → 18 min. (mid-2025) via AI-powered automation (ReliaQuest)
- Migrated to DragonForce April 2025 — tooling and affiliates persist under new brand

SCATTERED SPIDER

SOCIAL ENGINEERING

- English-speaking; native-language vishing bypasses training designed for foreign-accented attackers
- SIM swapping + help-desk impersonation to reset MFA — no malware needed for initial access
- MGM (Sep 2023): one help-desk call → $100M+ losses, 6TB stolen; Caesars paid $15M ransom
- **Increasingly using AI-generated voice tools for vishing** — lowers the skill bar for an already accessible attack
- 5 charged Nov 2024 (ages 20-23); UK retail attacks 2025 (M&S, Co-op, Harrods) — est. £440M in losses

Nation-State Adversary Profiles

CHINA

MOST CAPABLE

- Five-Year Plan targets 'new generation AI' — CCP aligns cyber ops to national plan
- Building own LLMs (Alibaba, Baidu) with heavy state direction
- AI used agentically to execute espionage — not just as advisory tool
- Volt Typhoon pre-positioned inside U.S. critical infrastructure IT environments, with concern about disruptive access
- Industrial-scale distillation attacks against U.S. AI models

RUSSIA

INFLUENCE + BATTLEFIELD AI

- Pioneered AI on battlefield — anti-drone operations
- AI-generated deepfakes for disinformation campaigns
- Uses criminal ransomware groups as cyber proxies
- OpenAI disruption (May 2024) of Russian covert influence ops
- Experimenting with AI to build and test model infrastructure
- DOJ Operation Doppelganger (Sep. 2024): dismantled Russian state-sponsored network using fake online personas and AI-generated content to spread political disinformation across the U.S.

IRAN

INFLUENCE + INFRASTRUCTURE

- Uses ChatGPT and others for research and content generation
- AI-generated synthetic media: fake missile/aircraft battle videos
- Influence ops: translated articles, webtags, headlines at scale
- Targeting critical infrastructure with AI-accelerated techniques
- OpenAI disrupted Iranian covert influence operation (May 2024)

NORTH KOREA

CRYPTO + FUNDING

- Crypto theft funds regime + AI development — 'easiest revenue source'
- Increasingly sophisticated laundering using AI tools
- AI-enhanced spearphishing of crypto platforms and exchanges
- Proceeds fund weapons programs and ongoing cyber capability build
- Targeting defense industry for IP to accelerate AI weapons R&D
- DPRK IT Worker Scheme: thousands of North Korean IT workers embedded in Western companies under false identities — generating regime revenue, stealing IP, and enabling insider access

TOPIC 1: THE THREAT LANDSCAPE — OT AND LEGACY SYSTEMS

Anthropic reports that Mythos finds 27-year-old zero-days.

OT systems run 27-year-old software.

Many cannot be patched at all.

Most Exposed Critical Infrastructure Sectors

Energy / Utilities

ICS/SCADA systems control the electrical grid, pipelines, and power distribution. Volt Typhoon (PRC) has been pre-positioned inside U.S. energy OT networks since at least 2023 — waiting. A Mythos-class zero-day converts that access into a destructive attack with no warning.

Water & Wastewater

Most underdefended U.S. critical infrastructure. Treatment PLCs run unpatched firmware from the 1990s. AI can compress the human-speed intervention window — a remote attack can alter chemical dosing or disable treatment before any operator can detect it.

Healthcare

Legacy medical devices — infusion pumps, imaging systems, ventilators — run embedded OS versions that cannot be updated without FDA re-certification. Hospitals cannot take them offline for patching. Mythos-class scanning could identify every one immediately.

Manufacturing

Industrial PLCs and CNC controllers often have no patch mechanism — vendor no longer exists or firmware is locked. A single Mythos-class zero-day chains directly to production shutdown, supply chain disruption, and potential physical equipment damage.

Transportation

Rail signaling, air traffic control legacy components, and port logistics platforms run 1990s software architectures. Salt Typhoon showed China's willingness to pre-position in U.S. transportation communications — OT systems are the logical next escalation.

TOPIC 2

AI Espionage — The New Crown Jewels and the Intelligence Gap

AI National Security Prosecutions: Models and Hardware

AI MODEL & TRADE SECRET THEFT

CONVICTED **U.S. v. Linwei (Leon) Ding | N.D. Cal. | Jan 2026**

First-ever AI economic espionage conviction. Former Google engineer stole 2,000+ pages of TPU chip designs, GPU system specs, and SmartNIC architecture — the hardware backbone for training frontier models. Intended to aid PRC AI supercomputer development and his own China-based ML startup. 7 counts economic espionage + 7 theft of trade secrets.

INDICTED **U.S. v. Song Wu | N.D. Ga. | Sep 2024**

Engineer at AVIC (Chinese state-owned aerospace & defense conglomerate) ran multi-year spearphishing campaign impersonating colleagues to steal aerospace source code from NASA, research universities, and private defense companies. Target: AI-adjacent aeronautics and sensor source code for PLA military modernization. 14 counts wire fraud, aggr. ID theft.

PLEA **U.S. v. Chenguang Gong | C.D. Cal. | 2025**

Dual U.S.-China citizen at defense contractor transferred 3,600+ files containing infrared sensor and missile component designs — including AI-enabled targeting systems — to personal devices. Repeatedly applied to PRC Talent Programs. Emphasized military potential in communications to Chinese handlers. Pleaded guilty to theft of trade secrets.

AI HARDWARE SMUGGLING (EXPORT CONTROLS)

GUILTY PLEA **Op. Gatekeeper / U.S. v. Hsu | S.D. Tex. | Oct 2025**

FIRST AI chip smuggling conviction. Alan Hao Hsu / Hao Global LLC exported $160M in Nvidia H100 and H200 GPUs to China via relabeling scheme ('SANDKYAN' fake brand). Chips seized. Co-conspirators Gong and Yuan separately charged. Guilty plea; sentencing pending.

CHARGED **U.S. v. Ho, Raymond, Li, Chen | M.D. Fla. | Nov 2025**

Four defendants (2 U.S. citizens, 2 PRC nationals) ran Florida front company to purchase and illegally export 400 Nvidia A100 GPUs + 10 HPE H100 supercomputers to China via Malaysia and Thailand. $4M+ in wire transfers from Chinese firms. Max exposure: 50 years per defendant.

INDICTED **U.S. v. Liaw, Chang, Sun (Super Micro) | S.D.N.Y. | Mar 2026**

Largest AI smuggling indictment: Super Micro co-founder Liaw and executives conspired to divert $2.5B in AI servers to Chinese customers 2024-2025. Staged thousands of dummy servers in U.S. warehouses to deceive compliance teams. Transshipped via Taiwan. Chang remains a fugitive.

Three Novel Threat Vectors

MODEL WEIGHT THEFT

Model weights — the numerical parameters encoding AI intelligence — can be copied and exfiltrated digitally without physical transfer, replicating full model capabilities without billions in training cost.

- Weights can be exfiltrated in a single cloud upload — no physical extraction required
- A frontier model's core capabilities may be replicated if weights and serving infrastructure are stolen — erasing years of American compute advantage
- Biden's AI Diffusion Rule (Jan. 2025) — only formal export control ever imposed on model weights — rescinded by Trump Admin (May 2025) days before taking effect
- Existing authorities may apply, but no AI-weight specific authority squarely addresses risk. H.R. 8283 (Apr 2026) targets distillation only; GP10 catch-all requires proving military end use — a high bar

DISTILLATION ATTACKS

Systematic API querying of U.S. AI models to extract and replicate their capabilities — without physical theft, often within ostensibly authorized usage, creating novel legal challenges.

- DeepSeek R1 (Jan. 2025): widely cited alleged large-scale distillation example — systematically queried U.S. frontier models to replicate capabilities at a fraction of the training cost
- Anthropic (Feb. 2026): DeepSeek, Moonshot AI, and MiniMax used 24,000+ fraudulent accounts to generate 16M+ exchanges with Claude, extracting reasoning and agentic behaviors
- OpenAI (Feb. 2026): letter to House China Select Committee citing evidence of ongoing DeepSeek distillation via new, obfuscated methods bypassing access controls
- WH OSTP Memo NSTM-4 (Apr. 2026): accuses China of "deliberate, industrial-scale campaigns"; directs intel sharing with AI companies.

AI AS THE ATTACKER

Anthropic detected a Chinese state-sponsored group using agentic AI not as an advisory tool — but to autonomously execute the cyberattack itself

- AI selected targets, harvested credentials, decided what to exfiltrate, and executed lateral movement autonomously — no human operator directing each step (Sept. 2025)
- Attackers manipulated Anthropic's own Claude Code as part of the intrusion campaign (Nov. 2025)
- Traditional detection cannot match AI enabled exfiltration (speed v. precision)
- Same agentic capabilities Mythos demonstrated for defense may be operationalized by adversaries who have acquired comparable tools
- Prompt injection: adversaries embed adversarial instructions in data an AI agent processes — hijacking its actions mid-operation, redirecting tasks, or exfiltrating session context; most dangerous precisely when agentic AI has the broadest network access

AI Threats & the Public-Private Intelligence Gap

WHY SHARING IS HARDER NOW

Adversaries Are Targeting the Sharing Mechanisms Themselves

- AI-enabled attacks move faster than legal and regulatory reporting timelines — by the time a company can share threat intel, the vector has already mutated
- Distillation, model theft, and prompt injection attacks — companies may not know they are victims until too late
- CISA 2015 liability shield expires Sept. 2026 — without it, voluntary sharing exposes companies to civil and regulatory liability, chilling the information flow law enforcement depends on
- AI agents with broad system access complicate attribution — companies can't always distinguish an insider threat from an AI-enabled external actor mimicking legitimate user behavior

WHAT LAW ENFORCEMENT NEEDS

The Gap Between What Companies Know and What Agencies Can Act On

- Early, proactive notification of intrusion techniques — before a breach spreads — requires trust relationships built before a crisis, not during one (see: Operation Wintershield)
- Private sector will be best positioned to analyze data and identify actionable leads – requires developing relationships with LE beforehand
- AI-generated evidence is harder to preserve and authenticate — forensic standards for LLM-assisted attacks remain undeveloped, creating proof challenges in prosecutions
- WH OSTP Memo NSTM-4 (Apr. 2026): directs intel sharing with AI companies — but no reciprocal mandate for companies to share back with LE; gap persists

WHAT BETTER LOOKS LIKE

Closing the Loop: Models for Better Public-Private Sharing

- Operation Wintershield model: FBI proactively building company relationships before incidents; pre-negotiated SOPs for AI-specific threat intel sharing should become standard
- CIRCIA-like rule could consider AI-specific incident categories — including model theft, distillation attacks, and AI-enabled lateral movement — with dedicated LE notification pathways
- CISA 2015 liability shield reauthorization is essential
- AI used to attack AI companies means the best source of threat intelligence is also a victim — structured sharing agreements, not ad-hoc notifications, may be best
- WIMWIG Act (reauthorize CISA until 2035): AI-specific information sharing provisions are the right direction — need to close the cycle between private discovery and public defense

TOPIC 3

Mythos-Class Tools — When AI Becomes the Most Dangerous Vulnerability

TOPIC 3: MYTHOS — PROJECT GLASSWING, COMPETITION & THE POLITICAL DISPUTE

WHAT IS CLAUDE MYTHOS PREVIEW?

- Per Anthropic, Mythos is first AI model to autonomously find AND exploit thousands of zero-day vulnerabilities — in every major OS, every major browser; 27-yr-old OpenBSD flaw, 16-yr-old FFmpeg flaw
  - Completed AISI's 32-step corporate network attack simulation; solved 73% of expert-level CTF problems — next-best model averaged 16 steps
  - Chains multiple vulnerabilities into working exploits without human direction — 20-gadget ROP chain against FreeBSD; 4-vulnerability browser sandbox escape
  - Near-0% autonomous exploit rate for prior model (Opus 4.6); Mythos represents a qualitative threshold never before crossed
- Model deemed too dangerous for general public release — access restricted to Project Glasswing partners only, but early reporting alleges attempted/unauthorized access via 3rd party environment; Anthropic investigating

PROJECT GLASSWING — GOVERNMENT & SECTOR RESPONSE

- 12 Launch Partners: Anthropic · AWS · Apple · Broadcom · Cisco · CrowdStrike · Google · JPMorganChase · Linux Foundation · Microsoft · NVIDIA · Palo Alto Networks — plus ~40 additional critical infrastructure orgs
- Same day Glasswing launched, Treasury and the Fed called an emergency meeting with the major banks not yet inside it — urging them to harden defenses before comparable capabilities spread
- Other agencies' use: CISA v. NSA; WH restrictions on private sector v. allowing federal agencies

THE COMPETITION — MYTHOS IS NOT ALONE FOR LONG

- Equivalent capabilities broadly available within 6-12 months — not just by companies in the US
  - Ex: OpenAI GPT-5.4-Cyber and Google Big Sleep already possess comparable vulnerability-discovery capabilities (Bain, Apr 2026)
  - Ex: China's Qihoo 360: 'Multi-Agent Collaborative Vulnerability Discovery System' claimed to find 1,000+ vulns using AI — including CVE-2026-32190, allegedly undetected 8 years (SecurityWeek, Apr 2026)
- Chinese law requires private firms to report vulnerabilities to government BEFORE public disclosure — effectively channeling AI-discovered zero-days to PLA and MSS
- AISLE: independent AI cybersecurity firm has been running autonomous discovery since mid-2025 — 180+ externally validated CVEs across 30+ infrastructure projects

IS IT A "SUPPLY CHAIN RISK"? THE LIVE LEGAL QUESTION

- Feb. 27: Sec. Hegseth designated Anthropic a 'supply chain risk' — President ordered all federal agencies to immediately cease using Anthropic tools. Root cause: Anthropic held two red lines — no autonomous weapons, no domestic mass surveillance
- Two court cases:
  - Mar. 26: NDCA issued temporary injunction blocking the designation as "classic First Amendment retaliation"; government intends to appeal
  - Apr. 8: DC Cir. denied Anthropic's request for a stay, but expedited review
- Apr. 17: Anthropic CEO Dario Amodei met WH Chief of Staff
- Now: WH now drafting potential AI executive order — core question unresolved: who decides the limits on AI use for national security purposes?

TOPIC 4

Legal & Compliance — What Your Organization Needs to Consider

Should Your Organization Seek Access to Mythos-Class Tools?

MODEL GOVERNANCE / INTERNAL CONTROLS

Organizations deploying high-risk AI capabilities should implement governance controls analogous to dual-use technology management — e.g., strict access controls, defined use-case limitations, audit logging, and escalation protocols.

CFAA ANALYSIS

Use of autonomous AI tools in internal testing raises unresolved questions under the CFAA's "exceeds authorized access" prong—particularly where agentic systems operate across systems or data beyond the original scope of authorization.

THIRD-PARTY EXPOSURE

When AI tools identify vulnerabilities in third-party or embedded vendor code, even limited exploitation during testing may create liability under contract, tort, or anti-hacking statutes depending on scope and authorization.

Best practice: coordinated disclosure to affected vendors before proof-of-concept use.

PRIVILEGE PROTECTION

Organizations should structure AI-enabled testing to support attorney-client privilege and work-product protection (e.g., counsel-directed testing, documented purpose, controlled distribution of findings).

CYBER INSURANCE GAP

Coverage may be uncertain depending on policy language, including exclusions for intentional acts, penetration testing activities, or use of unapproved tools.

AI-assisted testing that causes collateral effects may trigger coverage disputes or reporting obligations.

EU AI ACT

The EU AI Act's GPAI systemic-risk obligations become enforceable August 2, 2026; questions on scope and extraterritorial application are evolving and unsettled.

EXPORT CONTROL RISK

AI systems capable of autonomous vulnerability discovery and exploitation may raise export-control questions analogous to intrusion software and strong cryptography under the EAR.

While no formal classification currently exists, agencies are actively evaluating whether such capabilities fall within existing controls or require new rulemaking.

TOPIC 5

Defensive Posture — Updating Your Strategy for the Mythos Era

The Four Pillars — Now Stress-Tested by AI

Stop Initial Access

- Phishing-resistant MFA
- Accelerated patching
- Anomalous login monitoring

⚠ Mythos-Era Update

AI-assisted social engineering can undermine MFA, via real-time voice deepfakes in fatigue attacks. Controls must be hardened beyond standard phishing-resistant MFA.

Detect Lateral Movement

- Behavioral baselines
- Least-privilege architecture
- Endpoint & network telemetry

⚠ Mythos-Era Update

AI tools that simulate legitimate user activity are evading behavioral baselines. Zero trust is now mandatory, not aspirational.

Disrupt Exfiltration

- Behavior-based detection
- Outbound traffic monitoring
- Immutable backups

⚠ Mythos-Era Update

After months of AI-enabled dwell time, immutable backups are necessary but insufficient. Assume the adversary already knows your network better than you do.

Build Resilience

- Tabletop exercises
- Anti-ransomware platform
- Incident response planning

⚠ Mythos-Era Update

Establish a dedicated AI threat war room. Treat Glasswing patch outputs as emergency remediation priorities — these are not optional.

Operation Wintershield (FBI): Build your FBI relationship BEFORE a breach, not during one. Risk-based vulnerability management + pre-established response plans.

TOPIC 6

Regulatory Landscape — CIRCIA, CISA 2015, the SEC, and the AI Reporting Gaps

What Companies Must Report — and What the Law Doesn't Yet Cover

CIRCIA: THE STATUTE AND IMPLEMENTING RULE

- NPRM (Apr. 2024) proposed:
  - 72-hour reporting for covered cyber incidents; 24-hour reporting for ransomware payments
  - ~316,000 entities across 16 critical infrastructure sectors covered (may narrow)
- Target date of May 2026 for Final Rule is in flux
- CISA 2015 liability shield is currently in effect — extended through Sep 30, 2026 only after a 6-week lapse in late 2025; long-term reauthorization still unresolved in Senate
  - WIMWIG Act (H.R. 5079, 10-yr reauthorization with AI-specific provisions) cleared House Homeland Security Committee 25-0 (Sep. 2025); full House and Senate passage pending

AI-SPECIFIC GAPS NOT YET ADDRESSED

- AI-discovered zero-day not yet exploited — is it a reportable 'covered cyber incident'?
- Mythos-class scanning of embedded third-party code — does finding trigger reporting to that vendor?
- When an AI model itself is poisoned, inverted, or its weights stolen — what are the obligations?
- AI agent with broad network access is compromised and used for lateral movement — CIRCIA covers this?
- SEC material disclosure: what must public companies say about AI security posture in 10-K / 8-K?
- EU AI Act (Aug. 2, 2026): systemic-risk model enforcement — extraterritorial reach still unsettled

SEC 2026 Priorities emphasize AI, Cybersecurity, and Information-Security Risks; Crypto is less prominent. AI Washing is now an active enforcement concern

The Regulatory Calendar for AI & Cybersecurity (2026)

MAY 2026 (??) — IMMEDIATE

CIRCIA Final Rule

Proposed: Mandatory 72-hr incident + 24-hr ransomware payment reporting for ~316,000 critical infrastructure entities. Build compliance infrastructure now — don't wait for ink to dry.

NOW — WATCH

CISA 2015 Reauthorization

Liability shield lapsed Sept. 30, 2025 for 6 weeks, then extended through Sept. 30, 2026 via Consolidated Appropriations Act. Long-term reauthorization (WIMWIG Act) remains stalled in Senate. Companies should prepare for another potential lapse.

AUG 2, 2026 — UPCOMING

EU AI Act Enforcement

Commission power to fine providers of GPAI models with systemic risk. Up to 3% annual worldwide turnover. Extraterritorial reach for models used in EU still unsettled.

NOW — ACTIVE

SEC AI/Cyber Examinations

2026 priorities: cybersecurity and AI displace crypto. AI washing enforcement active. Material disclosure obligations for AI security posture in public company filings.

EMERGING — WATCH

Export Controls on AI Offensive Tools

Some legal experts argue AI models capable of autonomously exploiting zero-days may meet the functional definition of offensive cyber tools — though no formal USG guidance has classified them as such. EAR/ITAR frameworks for strong cryptography are the historical analogue; State/Commerce expected to develop AI-specific guidance.

ONGOING — ONGOING

DOJ/FBI Enforcement

Active priorities: PRC cyber intrusions, AI trade secret theft, export control violations for AI chips, FARA enforcement for AI-assisted foreign influence ops.

TOPIC 7

The Governance Frontier — What are the Rules Going Forward?

TOPIC 7: THE GOVERNANCE FRONTIER – WHAT ARE THE RULES GOING FORWARD?

Who Gets to Set the Limits? The DOJ / Pentagon / Anthropic Dispute

Export Controls for AI Offensive Tools

The Agentic AI Governance Gap and Removing the Human from the Loop?

How Can We Solve the Public-Private Intel Gap

By working together —
government, industry, and the private bar
—

we can protect our nation, our citizens, and our ideas.

Eun Young Choi Arnold & Porter | eunyoung.choi@arnoldporter.com

Cynthia Kaiser Halcyon.ai | linkedin.com/in/cynthia-kaiser-cyber

Sean Newell DOJ NatSec Cyber | www.justice.gov/nsd

Privacy + Security Forum | The Algorithm Arms Race | May 7, 2026