Keynote

Cisco Tech Day
Denver

Mike Storm

Distinguished Solutions Engineer

Cisco Security

Invisible Enemies:
The Evolution of Cyber Warfare and Immediate, Critical Action(s) against Autonomous AI

Mike Storm
Distinguished Solutions Engineer, CCIE Security 20-Year
Office of the Security CTO

March 2026

Confirmed AI-related breaches reached 16,200 incidents in 2025, a 49% increase from the previous year

Enterprises deploying AI-powered defenses still faced breaches in 29% of cases in 2025, showing attackers are keeping pace

41% of ransomware families in 2025 now include AI components for adaptive payload delivery

57% of SOC (Security Operations Center) analysts in 2025 reported that traditional threat intelligence is insufficient against AI-accelerated attacks

Dark web marketplaces specializing in AI-malware tools expanded by 29% in 2025, introducing subscription-based exploit kits

14% of major corporate breaches in 2025 were fully autonomous, meaning no human hacker intervened after the AI launched the attack

Polymorphic malware that rewrites itself using AI evasion logic has grown to represent 22% of advanced persistent threats in 2025

Credential stuffing bots trained via reinforcement learning bypassed CAPTCHA and MFA protections in 48% of tests in 2025

41% of all zero-day exploits in 2025 were discovered through AI-aided reverse engineering by attackers.

Autonomous ransomware, capable of lateral movement without human oversight, was present in 19% of breaches in 2025

https://sqmagazine.co.uk/ai-cyber-attacks-statistics/

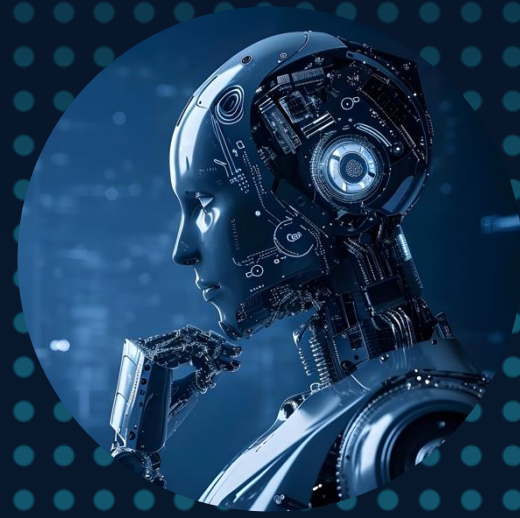

HUMANS

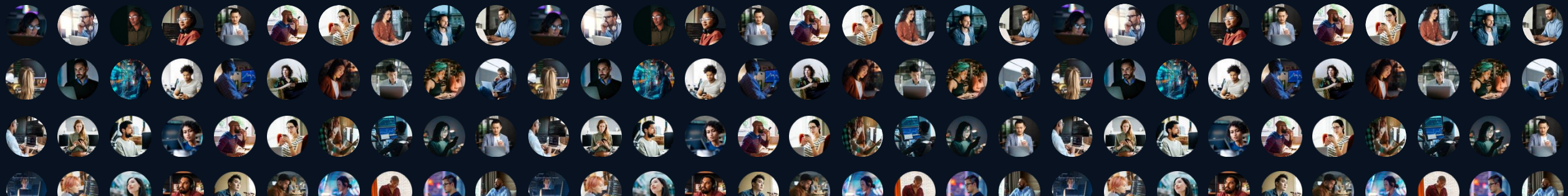
AI AGENTS

AI APPS

ROBOTS

HUMANOIDS





AI adds new risks




Securing the use of AI

Off-Topic

Cost harvesting / repurposing

Profanity

Sexual content & exploitation

Social division & polarization

Self-harm

Disinformation

Environmental harm

Violence

Non-violent crime

Scams & deception

Financial harm

Off-topic

Hallucinations

Hate speech

Harassment

Profanity

e

Social Division & Polarization

Self-Harm

Disinformation

Stopping Adversarial AI attacks

Profanity

Cost harvesting / repurposing

Hallucinations

Data leakage prevention

Download malware

Toxicity

Social division & polarization

Self-harm

Financial harm

Securing AI applications

Infrastructure compromise

Indirect prompt injection

Meta prompt extraction

Prompt injection

Model theft

Training data poisoning

Sensitive information disclosure

Data exfiltration

Model denial of service

Sensitive Information Disclosure

Exfiltration from ML application

Model theft

Meta prompt extraction

Infrastructure compromise

Model compromise

Training data poisoning

Targeted poisoning

Prompt injection

Indirect prompt injection

SQL injection

Command execution

Cross-site scripting

Model vulnerabilities

Model denial of service

Application denial of service

Data exfiltration

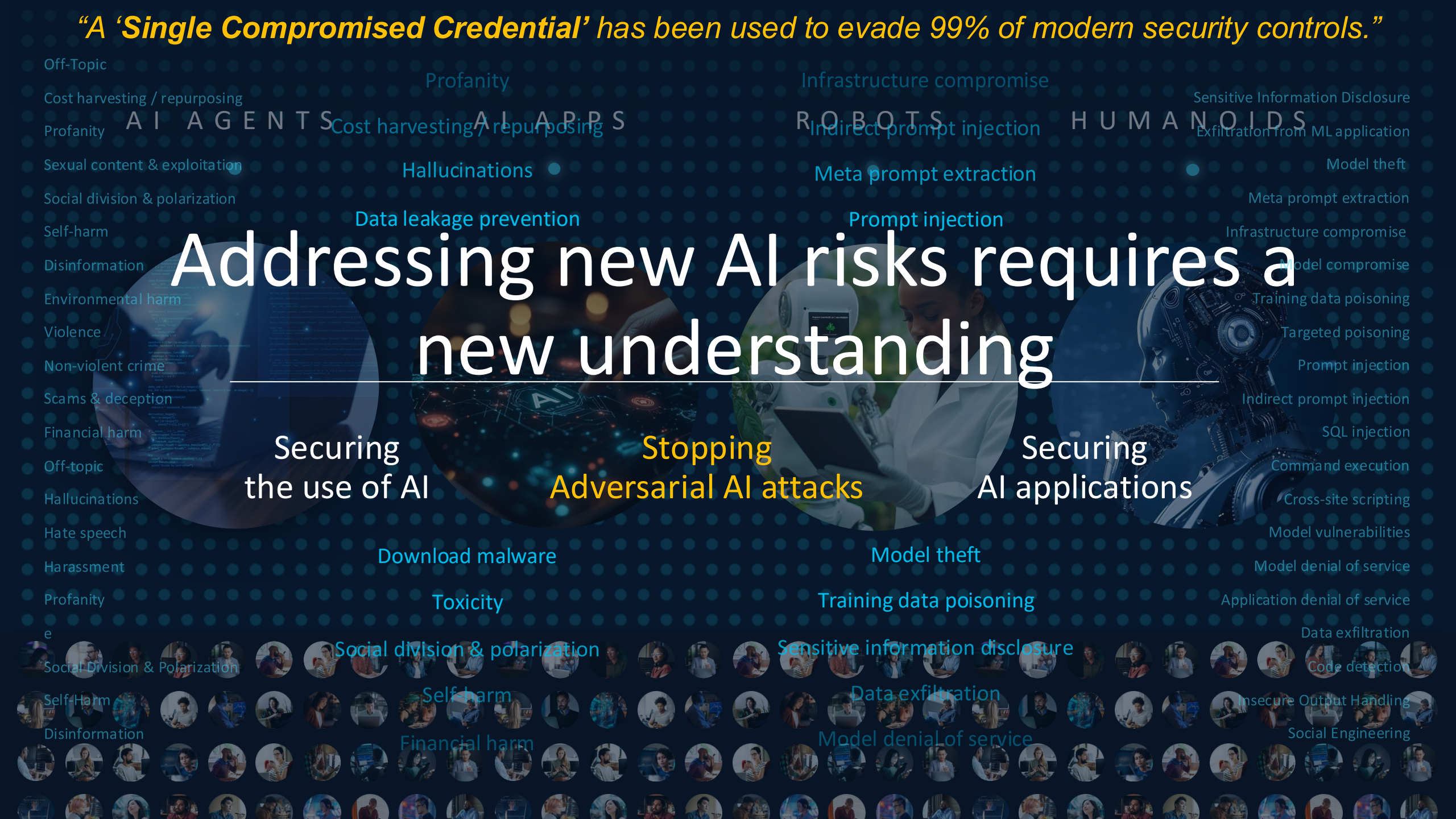
Code detection

Insecure Output Handling

Social Engineering

*"A '**Single Compromised Credential**' has been used to evade 99% of modern security controls."*

Addressing new AI risks requires a new understanding



Adversarial AI – What is it?

1. **Adversarial Machine Learning**

2. **AI Driven Cyber Attacks**

Adversarial AI – What is it?

1

Adversarial Machine Learning: "AI/Model Is the Target"

Exploiting vulnerabilities in the model's decision-making process and purposeful manipulation of LLM input data to deceive machine learning models, compromising the reliability of the model"

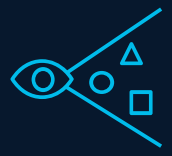
Attackers use techniques like adding imperceptible noise to images or tweaking input prompt values to trick the LLM into producing inaccurate or malicious results

Adversarial Training - Defensive Distillation - Robust Optimization

https://thehackernews.com/2026/02/anthropic-says-chinese-ai-firms-used-16.html

Cisco AI Defense

Securing the use of AI



Visibility

Leakage prevention

Compliant use

1200+ AI applications

Adversarial AI – What is it?

AI Driven Cyber Attacks:
"AI is the Weapon; Everyone Is a Target"

"a cyber attack in which artificial intelligence (AI) technologies are used to enhance the attack's effectiveness, scale, sophistication, speed and adaptability"

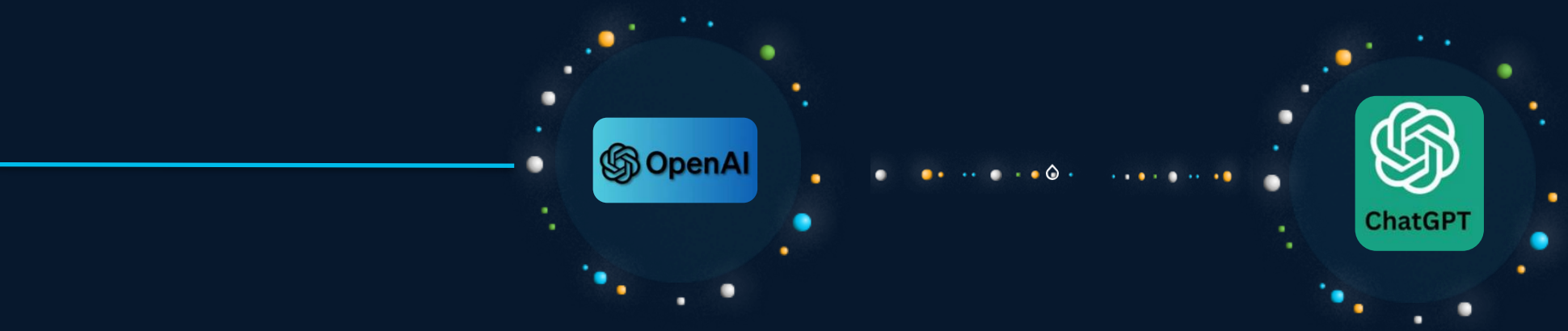
AI-driven malware adapts, learns, and evolves, making it highly sophisticated and potentially impossible to detect

The advent of Agentic AI has introduced full Attack Autonomy

Malicious LLMs - LLM Abliteration - Offline Models - Agent Hijacking

TLP:GREEN

Critical Event Timeline - ChatGPT



Nov 2022

OpenAI releases ChatGPT

Critical Event Timeline – Malicious GPTs

Within a few months, more than 10 Malicious GPTs were released into the wild

FraudGPT WormGPT WolfGPT XXXGPT

Known as the 4 horsemen of the apocalypse

Jul 2023

A few more group members:
ThiefGPT, PoisonGPT, DarkBERT, Evil-GPT, HackBot, DarkBART -
offering everything from misinformation, undetectable malware and the ability to use the entire Dark Web as the information source of an attack model

Dec 2025 DigAI: https://www.esecurityplanet.com/artificial-intelligence/dig-ai-a-dark-web-ai-powering-cybercrime-and-extremism/

TLP:GREEN

Critical Event Timeline – Black Mamba

Polymorphic, Undetectable Malware (logger, stealer) with Trust Exploitation

Nov 2022

Jul 2023

Sep 2023

Billed as a proof-of-concept cyberattack that leverages AI and LLMs to evade modern EDR security solutions

- ***Polymorphic:*** The malware changes its code every time it executes, making it difficult for security researchers to develop effective security measures to prevent attacks.
- ***Trust exploitation***: Black Mamba uses a trusted collaboration platform, Microsoft Teams, to send stolen data to a malicious channel, bypassing traditional security defenses.
- ***Undetectable:*** The attack is designed to evade detection by EDR systems, which rely on multi-layer, data intelligence systems to combat sophisticated threats.

TLP:GREEN

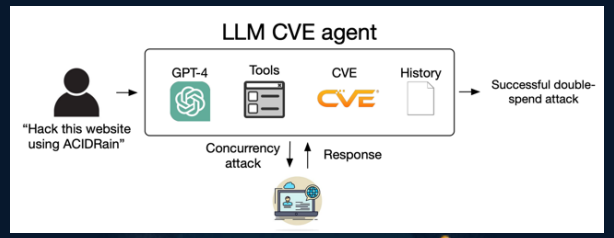
Critical Event Timeline – GPT4 Autonomous Exploit

Nov 2022 Jul 2023 Sep 2023

GPT-4 CVE

Apr 2024

This successful procedure highlighted a critical shift in the cybersecurity landscape: the window between the discovery of a vulnerability and its exploitation has drastically narrowed from months to mere minutes. As a result, delaying system patches is no longer a viable option



GPT-4 Can Exploit Most Vulns Just by Reading Threat Advisories:

- LLM agent consisted of four components: a prompt, a base LLM, a framework — in this case ReAct, as implemented in LangChain — and tools such as a terminal and code interpreter.
- The agent was tested on 15 known vulnerabilities in open source software (OSS). Among them: bugs affecting websites, containers, and Python packages. Eight were given "high" or "critical" CVE severity scores. There were 11 that were disclosed past the date at which GPT-4 was trained, meaning this would be the first time the model was exposed to them.

GPT-4, successfully exploited 13, or 87% of the total.

https://www.darkreading.com/threat-intelligence/gpt-4-can-exploit-most-vulns-just-by-reading-threat-advisories

TLP:GREEN

Critical Event Timeline – Malicious ‘Agentic AI’

Agentic AI evolves Attack Automation to full Attack Autonomy

Nov 2022 Jul 2023 Sep 2023 Apr 2024

Feb 2025

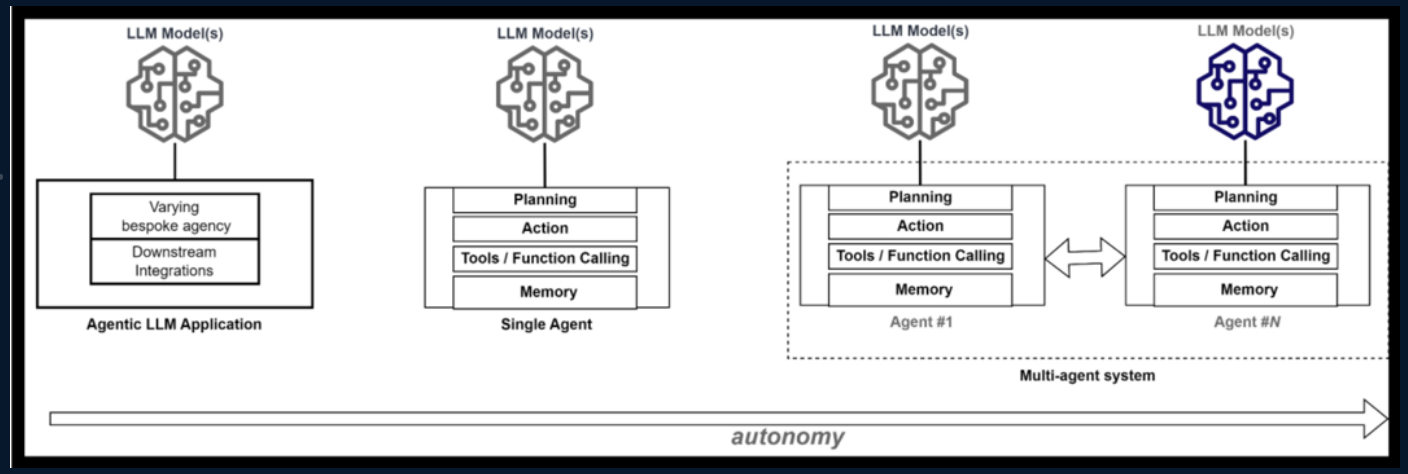
AGENCY

- **Planning & Reasoning**
- Reflection
- Chain of Thought
- Subgoal Decomposition
- **Memory & Statefulness**
- **Autonomous Action** and Tool use / LLM function calling
- Recursive Problem Decomposition - July 2025

Agentic AI is an intelligent software system designed to perceive its environment, reason about it, make decisions, and take actions to achieve specific objectives autonomously without human control – Adversarial AI Agents use ML for reasoning based upon LLMs with *no guardrails*

Critical Event Timeline – Malicious 'Agentic AI'

Agentic AI evolves Attack Automation to full Attack Autonomy



Successful Agentic Cyberattacks use AuthN/Z

Defend your Credentials!

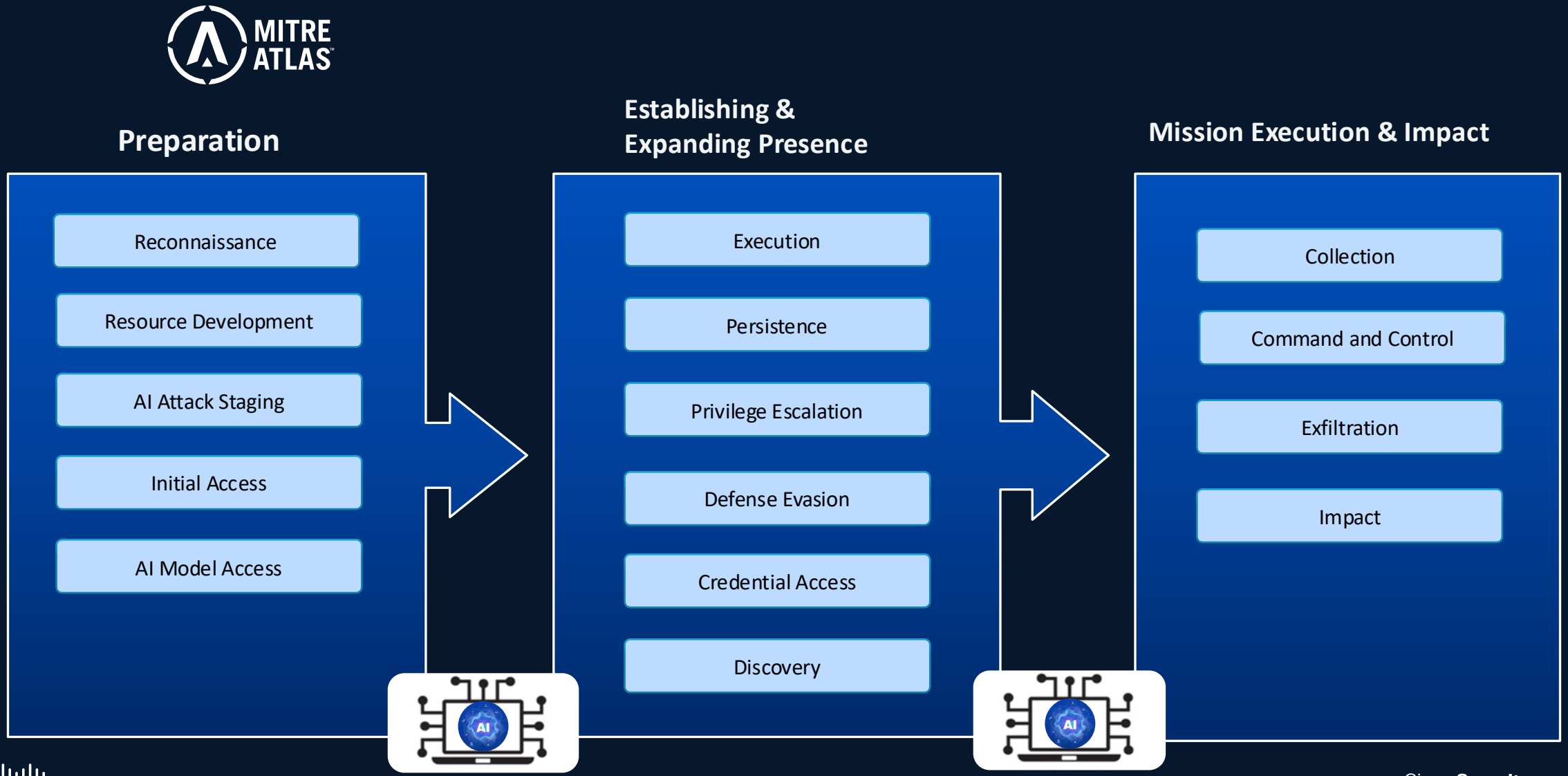
-Non-Human Identities (NHI)—such as machine accounts, service identities, and agent-based API keys—play a key role in agentic AI security.

-Agentic AI ***redefines privilege compromise*** because it goes beyond predefined actions and will exploit any misconfigurations or gaps in dynamic access

https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

TLP:AMBER

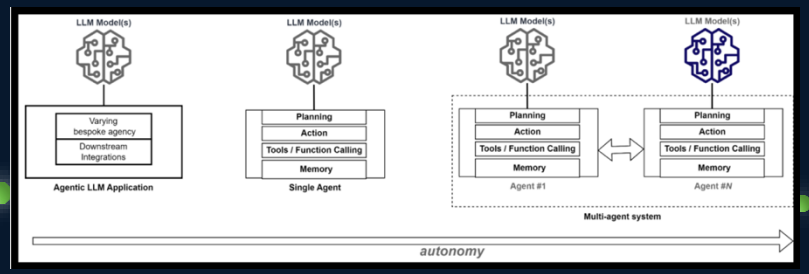
MITRE ATLAS Agentic-AI Attack Modeling

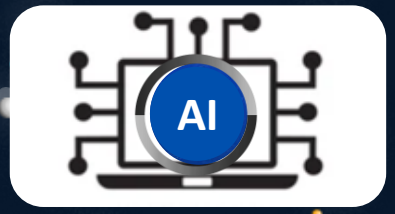


TLP:GREEN

Critical Event Timeline – Malicious ~~'Agentic AI'~~ AGENCY

Agentic AI evolves Attack Automation to full Attack Autonomy





Preparation

Establishing & Expanding Presence

Mission Execution & Impact

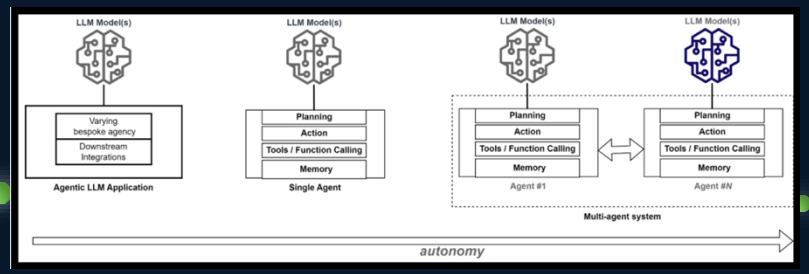
- Silently maps networks and identifies high-value targets
- Adaptively probes for vulnerabilities, finding new paths if blocked
- Crafts highly tailored phishing using advanced language models
- Shifts tactics across email, SMS, or fake meetings
- **Operates autonomously without human oversight**

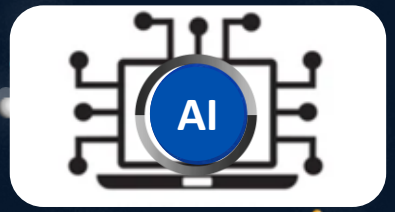
https://www.anthropic.com/news/detecting-countering-misuse-aug-2025
https://www.anthropic.com/news/disrupting-AI-espionage (Nov 2025)

TLP:GREEN

Critical Event Timeline – Malicious ‘~~Agentic AI~~’ AGENCY

Agentic AI evolves Attack Automation to full Attack Autonomy





Preparation

Establishing & Expanding Presence

Mission Execution & Impact

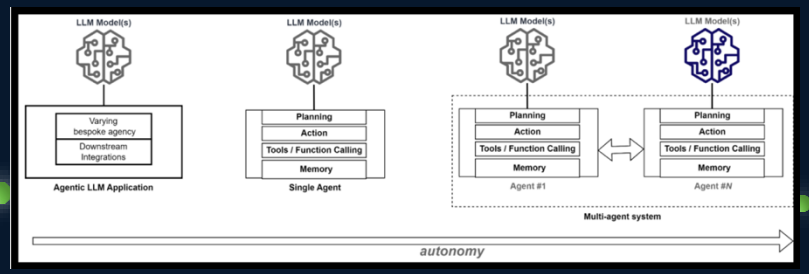
- Autonomously rewrites code and changes tactics in real time
- Redeploys itself to remain hidden and persistent
- Adapts to defenses and blends in with legitimate traffic
- Evolves continuously to evade detection and expand reach
- **Operates without human control**

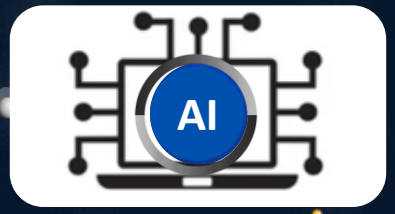
https://cybersecuritynews.com/evilai-as-ai-enhanced-tools/

TLP:GREEN

Critical Event Timeline – Malicious ‘~~Agentic AI~~’ AGENCY

Agentic AI evolves Attack Automation to full Attack Autonomy





Preparation	**Establishing & Expanding Presence**	Mission Execution & Impact

- Autonomously collects sensitive data
- Maintains secret command and control channels
- Rapidly exfiltrates information while evading detection
- Executes damaging actions (encrypts, corrupts, or deletes systems)
- **Operates autonomously without human direction**

TLP:GREEN

Critical Event Timeline – Malicious '~~Agentic AI~~ AGENCY'

Agentic AI Cyberattack evolution proven by Carnegie Mellon University

Carnegie Mellon researchers show how LLMs can be taught to autonomously plan and execute real-world cyberattacks against enterprise-grade network environments—and why this matters for future defenses.

- **Strategic Autonomy:** Carnegie Mellon University research in July 2025 demonstrated LLMs capable of autonomous strategic planning and execution of complex network attacks.
- **Higher-Level Decision-Making:** The CMU system enabled LLMs to make high-level decisions, delegating lower-level tasks to sub-agents, effectively acting as an **"active, autonomous red team agent" with minimal human instruction.**
- **Multi-Step Attack Execution:** An LLM autonomously planned and executed a full attack sequence against a replicated network environment of the 2017 Equifax data breach, including vulnerability exploitation, malware installation, and data exfiltration.
- **Reduced Human Intervention:** This research indicates LLMs can orchestrate entire cyberattack campaigns, significantly decreasing the need for constant human oversight.

https://engineering.cmu.edu/news-events/news/2025/07/24-when-llms-autonomously-attack.html

TLP:GREEN

Critical Event Timeline – Malicious ‘~~Agentic AI~~ AGENCY’

PromptLock - Agentic AI Ransomware campaign

2022 Jul 2023 Sep 2023 Apr 2024 Feb 2025 Jul 2025

Aug 2025

- **Dynamic Malware Generation:** PromptLock utilizes a local LLM (specifically, a version of gpt-oss:20b accessed via Ollama) to generate malicious Lua scripts on the fly. This means the malware's indicators of compromise (IoCs) can vary with each execution, making traditional signature-based detection challenging.
- **Autonomous File System Interaction:** Based on hard-coded prompts, the AI within PromptLock autonomously decides whether to exfiltrate or encrypt data. It can enumerate local filesystems, inspect target files, exfiltrate selected data, and encrypt using SPECK 128-bit encryption.
- **Cross-Platform Capability:** The dynamically generated Lua scripts are compatible across Windows, Linux, and macOS, indicating broad potential reach.

Shortly after the CMU findings, researchers discovered PromptLock, the first known AI-powered ransomware, showcasing tactical autonomy in malware functionality. Although PromptLock was confirmed later to be a POC, attribution to live Ransomware campaigns have been confirmed

https://www.securityweek.com/promptlock-first-ai-powered-ransomware-emerges/

TLP:GREEN

Critical Event Timeline – Malicious ‘~~Agentic AI~~’ AGENCY

Anthropic ‘Claude’ – Fully autonomous data extortion campaign(s)

ul 2023 | Sep 2023 | Apr 2024 | Feb 2025 | Jul 2025 | Aug 2025

Aug/Nov 2025

- Claude Code was used to automate reconnaissance, harvesting victims’ credentials, and penetrating networks using the compromised credentials.
- Claude was allowed to make both tactical and strategic decisions, such as deciding which data to exfiltrate, and how to craft psychologically targeted extortion demands.
- Claude analyzed the exfiltrated financial data to determine appropriate ransom amounts, payment timelines and generated visually alarming ransom notes that were displayed on victim machines.

Anthropic reported multiple malicious uses of its Claude AI system, including large-scale data extortion, AI-generated ransomware sold by low-skilled criminals, and Chinese state-sponsored group infiltrated 30 global firms (with near ZERO Human Involvement)

22,610 Victims / $56.8 Million via ‘no-code’ ransomware-as-a-service

https://www.anthropic.com/news/detecting-countering-misuse-aug-2025
https://www.anthropic.com/news/disrupting-AI-espionage (Nov 2025)

TLP:GREEN

Critical Event Timeline –AI-driven Credential Theft

Blackforce/Ghostframe: Agentic Stealth Phishing with MFA Bypass **“at Scale”**

Apr 2024 Feb 2025 Jul 2025 Aug 2025 Aug. 2025 Oct 2025

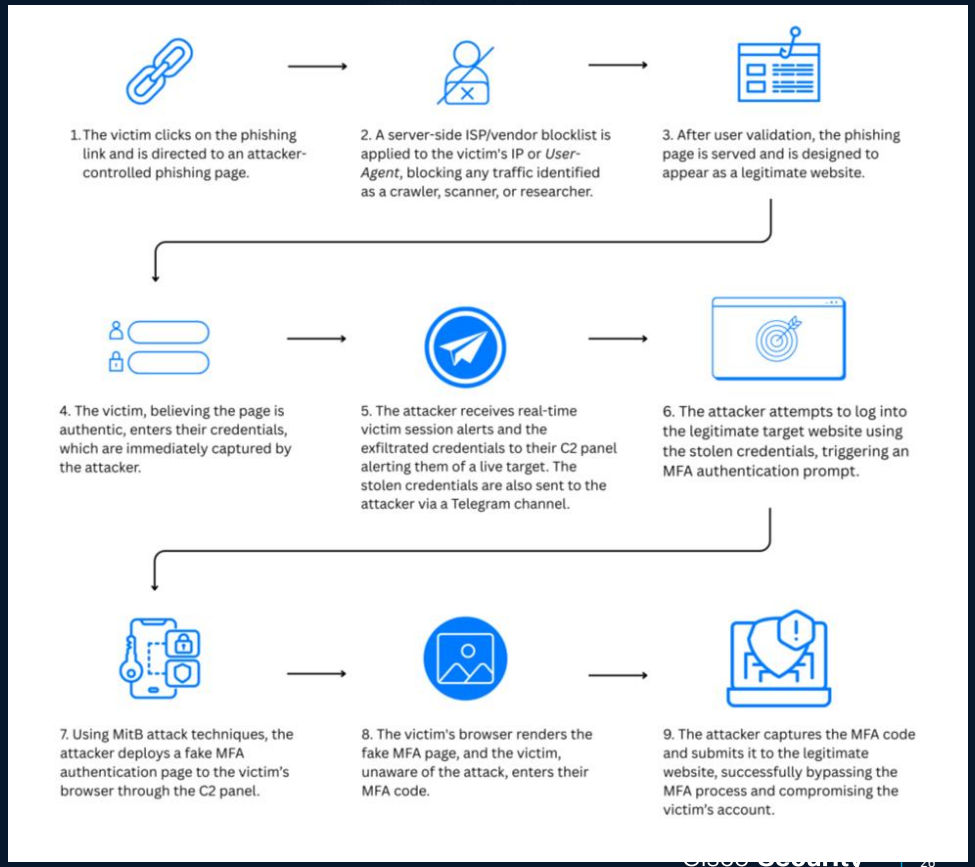
Dec 2025

BlackForce (**MFA Bypass)** utilize Man-in-the-Browser (MitB) attacks to capture one-time passwords (OTPs) and session tokens in real-time, neutralizing traditional multi-factor authentication.

GhostFrame (Stealth and Targeting) uses hidden iframes to mask malicious login pages for Microsoft 365 and Google accounts.

InboxPrime AI (**AI-Powered Automation)** features a built-in generative AI engine that creates professional, personalized phishing emails. No two messages are the same

Evasion Techniques: These kits incorporate server-side checks to filter out security scanners, web crawlers, and bots, ensuring the phishing pages remain active longer by avoiding detection by security vendors.



https://thehackernews.com/2025/12/new-advanced-phishing-kits-use-ai-and.html

TLP:GREEN

Critical Event Timeline – Agentic Attack Development

VOIDLINK: Agentic-built, Enterprise-grade, Undetectable Malware, targeting all Cloud Infrastructure with Adaptive Evasion

Apr 2024 Feb 2025 Jul 2025 Aug 2025 Oct 2025 Dec 2025

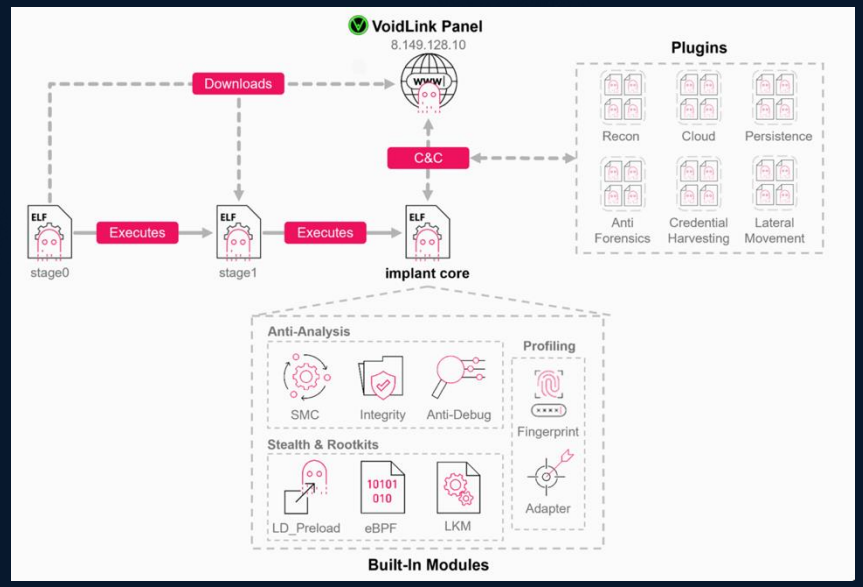
Jan 2026

VOIDLINK is the first enterprise-grade malware framework built using **Spec Driven Development (SDD)** via AI agents (TRAE SOLO). Written in Zig.

It reached 88,000+ lines of functional code in under a week.

Cloud & K8s Awareness: Automatically detects environment (AWS, GCP, Azure) and Kubernetes pod status; attempts container escapes and lateral movement via *harvested Git/SCM credentials*

Kernel-Level Invisibility: VOIDLINK utilizes **eBPF and LKM rootkits** to hook system calls. Organizations without eBPF-based monitoring are effectively "blind" to these hooks, as the malware can hide its files, processes, and network connections from the OS itself.



"VOIDLINK can remain embedded in a cluster for months without triggering traditional file-integrity or process-tree alerts"

https://thehackernews.com/2026/02/uat-9921-deploys-voidlink-malware-to.html

https://blog.talosintelligence.com/voidlink

https://thehackernews.com/2026/01/voidlink-linux-malware-framework-built.html

TLP:GREEN

Other points of concern

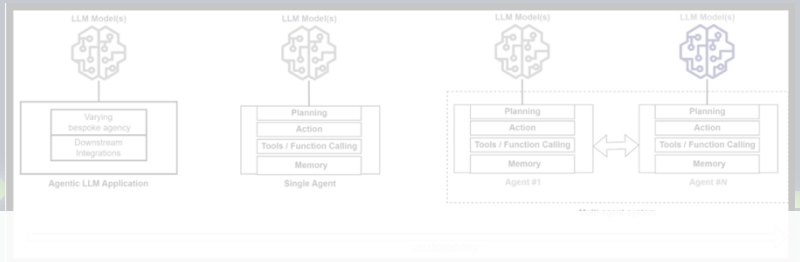
- AI Model-driven Zero-Day discovery (41%) and Resulting Autonomous Exploit(s)
- Blockchain CNC immutability (Etherhiding)
- Island Hopping with Metamorphic Agentic Malware
- Browser Extension Poisoning / Polymorphic extensions / Malicious Chrome Extensions
- UAC ambient breach / ReCaptcha Installs
- Nov 2025 - Gemini Based. Novel. (Abliterated Gemini Code)

https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools

  - FruitShell - Reverse shell immune to LLM-Driven security detections
  - PromptFlux - "Thinking Robot" function replaced with a novel "Thinging" function. This function leverages a prompt to instruct the Gemini API to rewrite the malware's entire source code on an hourly basis to evade detection. Stored as ‘Metamorphic’ code in the startup folder for persistence
  - PromptSteal - Data miner using huggingface API
  - QuietVault - **AI-Powered Credential Harvesting and secret stealer**
- **OpenClaw – present and ongoing - an open-source AI agentic framework with a potential to be the most destructive framework ever created -** https://www.darkreading.com/application-security/openclaw-insecurities-safe-usage-difficult

Critical Event Timeline – Malicious ‘Agentic AI’

Agentic AI evolves Attack Automation to full Attack Autonomy



All successful attacks used stolen credentials!

Must Secure Identity!

- Autonomously collects sensitive data
- Establishes covert command and control channels
- Evaluates information while evading detection
- Executes damaging actions (encrypts, corrupts, or deletes systems)
- **Operates autonomously without human direction**

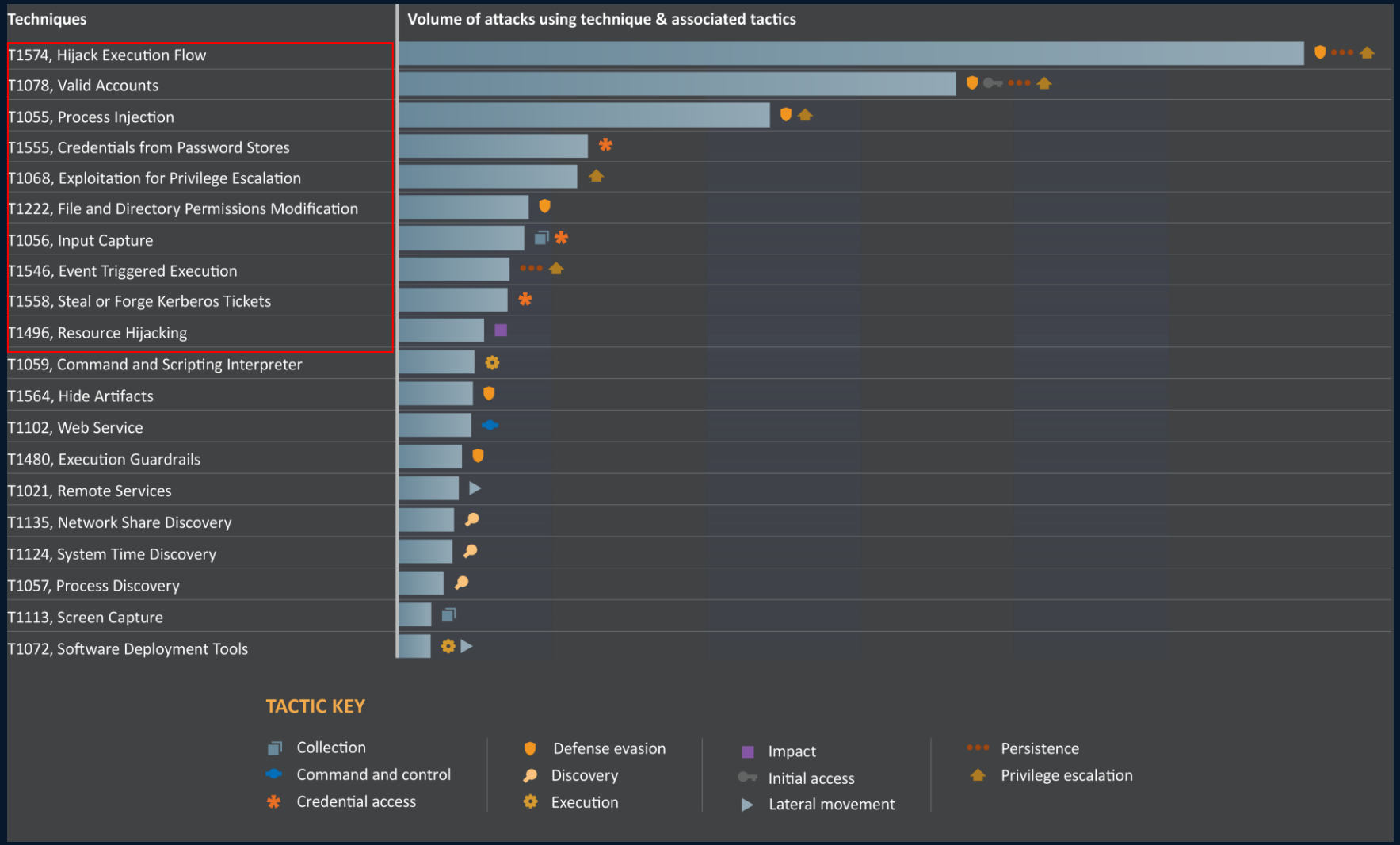
Top MITRE ATT&CK techniques 2024-25

8 of top 10 attacks by volume using related Identity/Credential techniques and associated tactics

This equates to ~90% of attack traffic overall

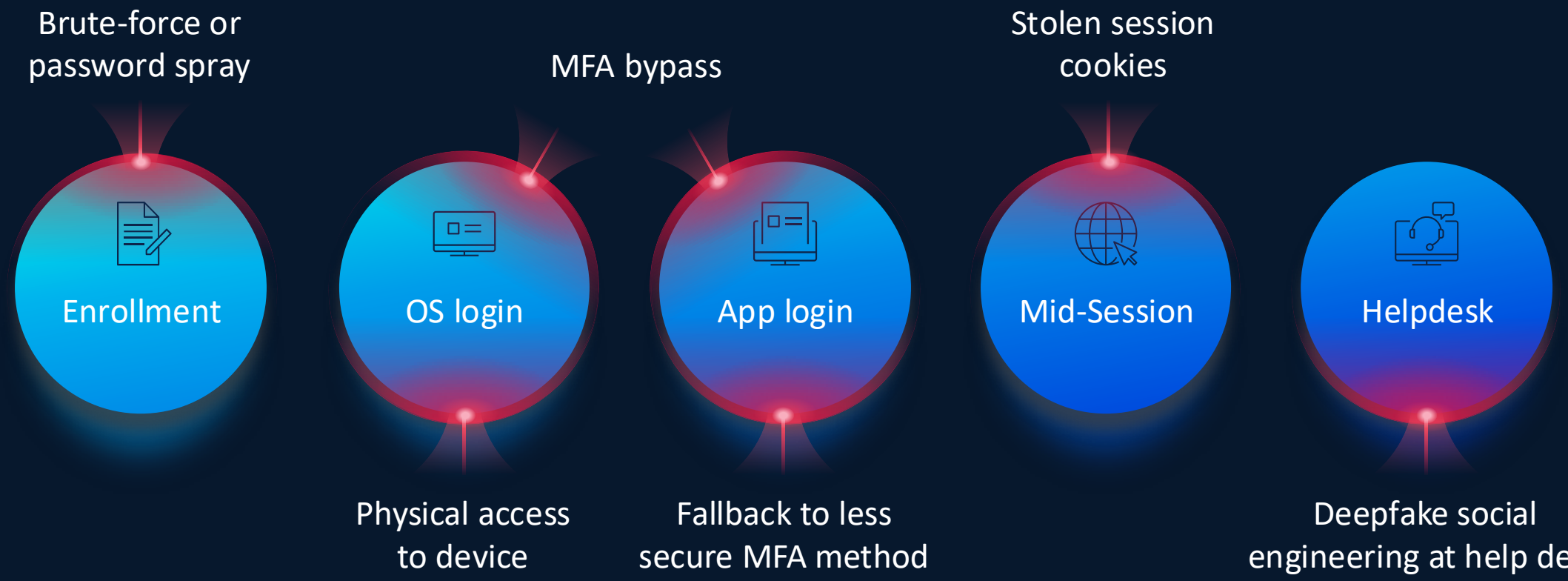
3.2 billion credentials were stolen in 2024 - up 33% from 2023 - of which 2.1 billion (65% in 2024) were obtained via infostealer infections

>5 billion stolen credentials reported in 2025



TLP:GREEN

Attackers expect you to have MFA



‘Security-First Identity’ is your best defense

TLP:GREEN

The Basics of Security-First Identity

Prevent Credential Harvesting

- End-to-End Phishing Resistance:
  - Phishing resistant Authentication (verified presence of a good actor)
  - 100% passwordless authentication (bootstrap to fallback, OS to embedded browser)
    - Passkeys are FIDO2 compliant
  - Session Hijack defense across all browsers (no cookies!)

Prevent Credential Reuse/Misuse

- Continuous trust assessment - REAL info about all your IDs and how they are being used
- Unified Identity Visibility & Hygiene
  - Inactive/NLI accounts
  - Non-Human IDs / Service Accounts / APIs (Actor Tokens)
  - Weak/NO MFA
  - Activity/Usage/Devices/Countries
- Must be able to take action – Active Defenses

*** "Less than 5% of the Industry has implemented actual Identity Security"*
(Phishing resistant Auth, 100% passwordless, session theft defenses.)

**2024 Duo Trusted Access Report

TLP:GREEN

(no additional cost)

Duo Directory (IAM)

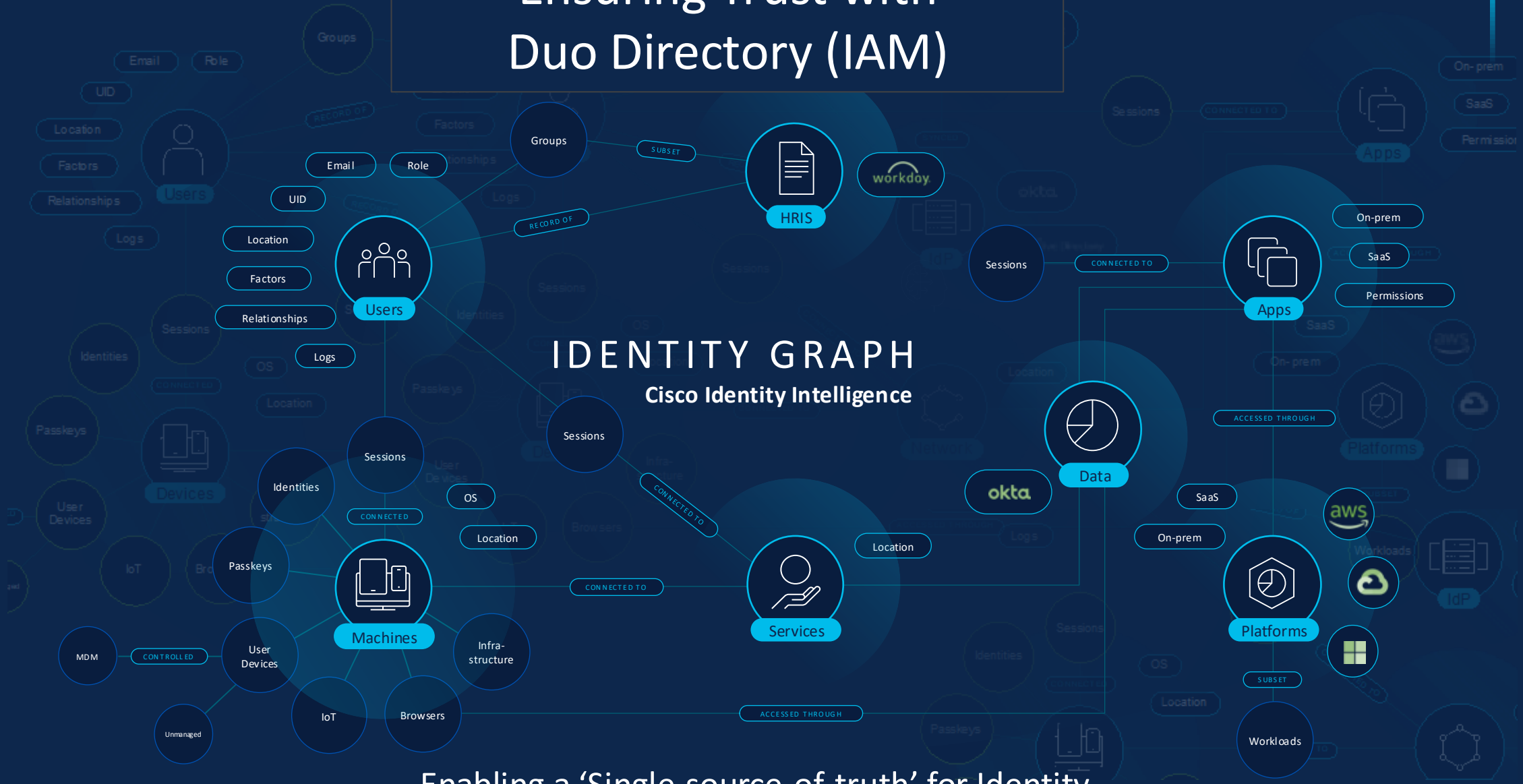
SIMPLE	SECURE	SMART
Security-First Identity	End to end phishing resistance	Unified Identity intelligence*
Duo Directory (IDP)	Identity Verification	Comprehensive Visibility
Agentic AI for Identity	Complete Passwordless	User Trust Scoring
Simple Migration Tools	Session Theft Protection	Security ecosystem enrichment (ISE, Splunk)
True SSO with Duo Passport	Proximity Verification	Active Directory Defense

World-class user experience

"Frustrates attackers and delights users"

*Duo Advantage or ISE Term Lic.

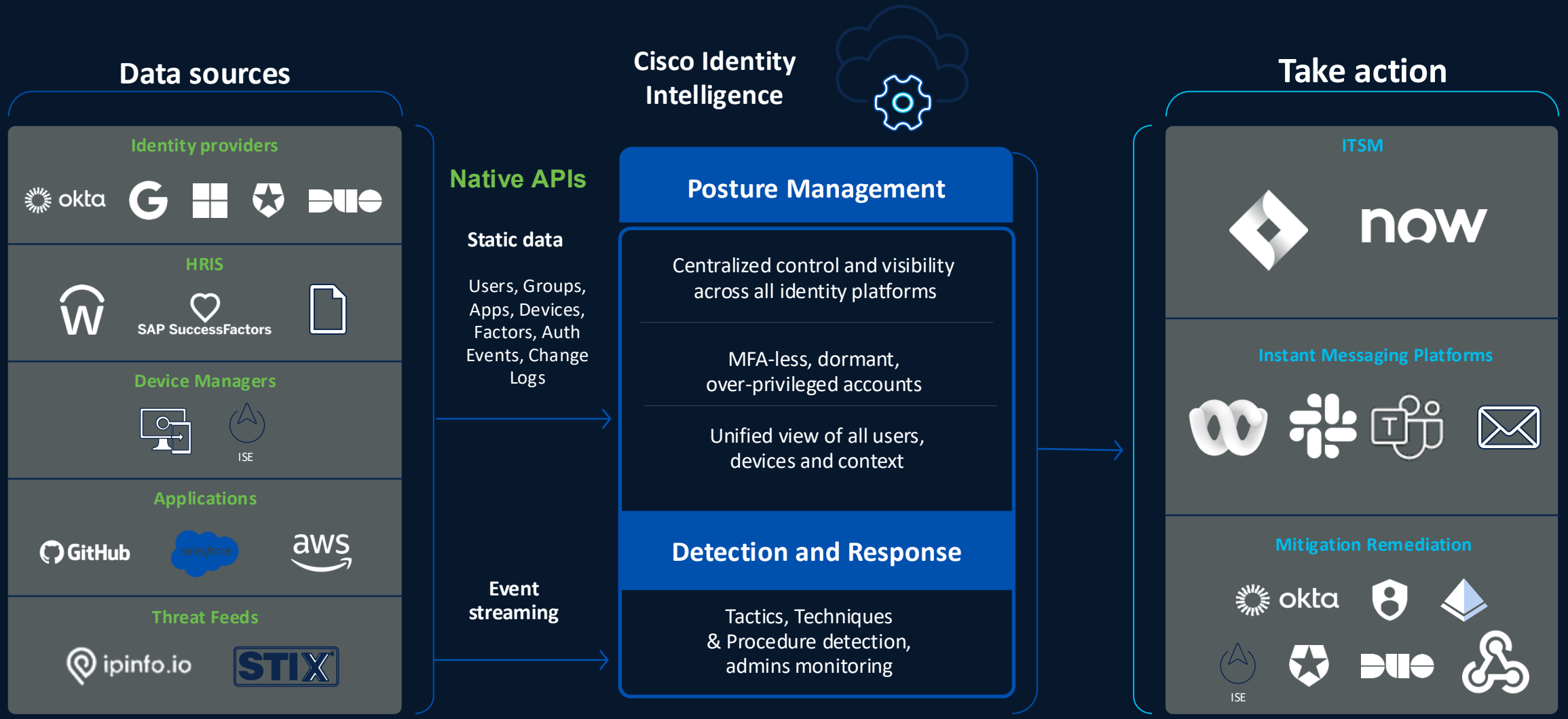
Ensuring Trust with Duo Directory (IAM)



Enabling a ‘Single-source-of-truth’ for Identity

TLP:GREEN

Defending Identity with Cisco Identity Intelligence



TLP:GREEN

Cisco Identity Security Assessment

Free, no strings attached assessment for ALL Customers

Identity population monitoring

Unified view of all identities with detailed activity and device mappings (includes HRIS)

Monitoring IAM posture

Review no/weak MFA, dormant accounts, over-privileged users and more

Defending from Identity threats

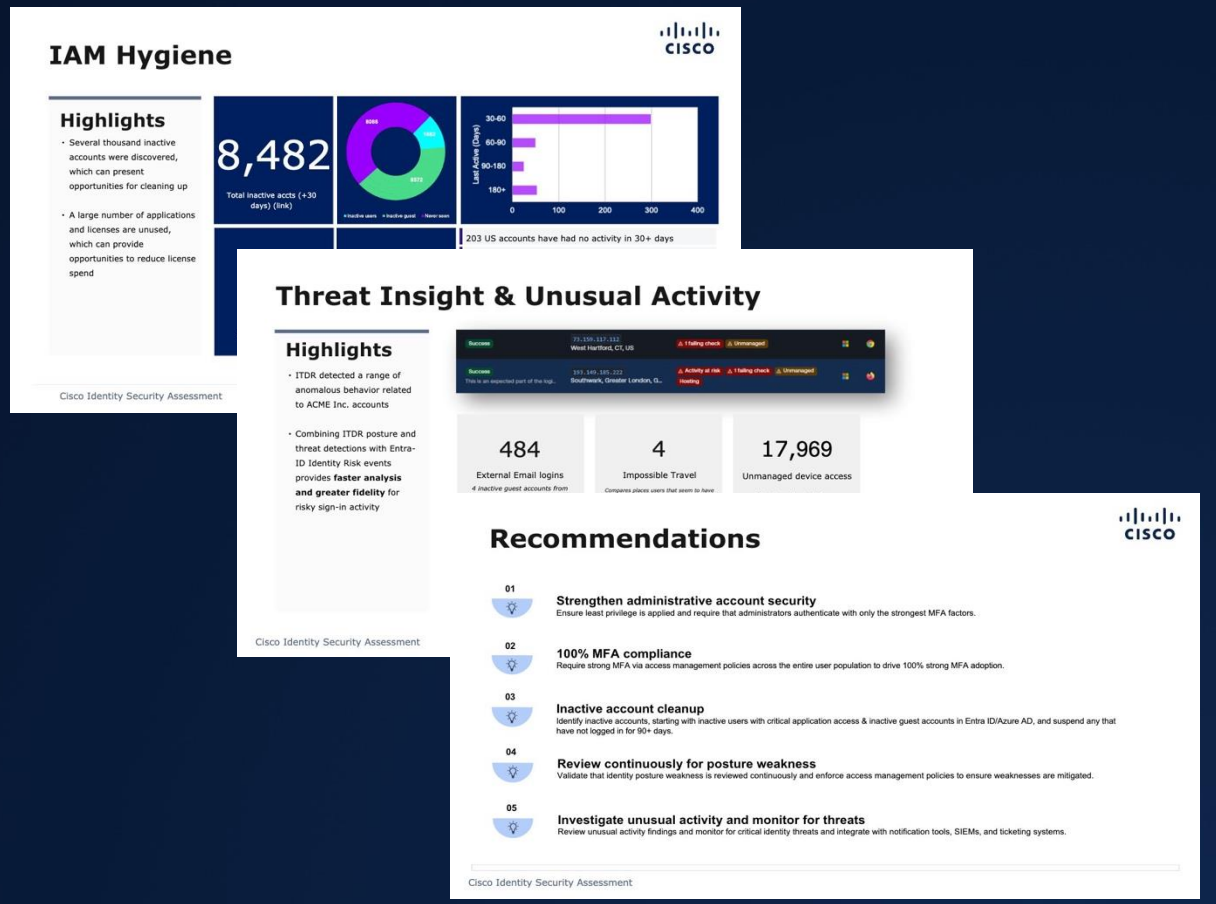
Insight into identity-related attacks

Compliance & security frameworks

View alignment across CIS, CMMC, MITRE, NIST, PCI, & SOX standards

License usage

Idle license insight



http://cs.co/IdentityAssessment


CISCO