AI AND CYBER RISK

Four warnings

Three institutions and one researcher who has just walked out of a frontier lab, inside three weeks.

1 **GOOGLE THREAT INTELLIGENCE GROUP** 9 SEP

forward leaning adversaries transition from basic prompting to agentic AI workflows

2 **NSA, CISA, FBI, ENERGY AND EPA** 19 AUG

This is not a theoretical risk ... it is an active threat

3 **THE FINANCIAL STABILITY BOARD** 28 AUG

may have the ability materially to alter the speed, scale and economics of cyber risk

4 **AI RESEARCHER, EX-ANTHROPIC** 9 SEP

will soon be superhuman systems that can hack anything

Amar Harolikar September 2026 tigzig.com

WARNING ONE · GOOGLE THREAT INTELLIGENCE

The attackers stopped typing and started delegating

Google Threat Intelligence Group runs a tracker on how attackers use AI. Its 9 September report is built on **frontline Mandiant incident response engagements, global threat actor tracking and live platform defences**.

> *GTIG has observed forward leaning adversaries transition from basic prompting to agentic AI workflows and AI-enabled automation. In these operations, human-in-the-loop latency is dramatically reduced, compressing the traditional window for defenders to respond.*
>
> **Google Threat Intelligence Group** · 9 September 2026

Under six hours

In the second quarter GTIG watched attackers compromise a cloud resource, then plan, build and execute a mass credential harvesting campaign, start to finish.

Four months earlier the same group reported the first zero day it believes was developed with AI, built for a mass exploitation event.

SOURCES GTIG, From Prompting to Autonomy, 9 Sep 2026 · GTIG, 12 May 2026

WARNING ONE, CONTINUED

And what they are going after

Four of the trends GTIG names for the second quarter of 2026.

Multi-agent frameworks

adversaries are deploying frameworks that autonomously manage scanning pipelines, resolve operational errors, and execute credential harvesting at scale

The software supply chain

one financially motivated actor has run large scale compromises of PyPI, npm and Docker Hub since March, and GTIG believes AI assisted coding contributed to the big supply chain compromises of the past eighteen months

AI itself as the target

adversaries are going after proprietary models, source code, prompts and API credentials, across healthcare, government and media

Somebody else's compute

stealing developer credentials and hijacking enterprise cloud accounts to run their own high performance workloads

GTIG names the actors it is describing, dates the activity, and says the report is grounded in its own incident response engagements. It is linked below.

SOURCES GTIG AI Threat Tracker, 9 September 2026

WARNING TWO · US CYBER DEFENCE

Public information, turned into working code

On 19 August the NSA, CISA, the FBI, the Department of Energy and the EPA issued a joint advisory. Attackers are using AI assistance to generate exploitation scripts from **publicly available information** about these controllers, built on an open-source library, and aiming them at internet exposed devices.

> *Using AI to generate exploitation scripts represents an evolution in threat actor capabilities, dramatically reducing the technical expertise and time required to develop working ICS exploitation scripts and malicious tools.*
>
> **NSA, CISA, FBI, Department of Energy and EPA** · advisory AA26-231A, 19 August 2026

The sectors they name are critical manufacturing, energy, water and wastewater, chemicals, food and agriculture, and commercial facilities. Their own words are that this **is not a theoretical risk** and that it is **an active threat**.

The blast radius is not one vendor and not one country. The advisory says the targeting is broader than the controllers it names, and that every owner and operator should apply the mitigations. Public information plus a model does not stop at one product line or one country.

SOURCES Joint advisory AA26-231A, 19 August 2026

WARNING THREE · FINANCIAL STABILITY BOARD

The one that reaches the financial system

On 28 August the chair of the Financial Stability Board wrote to G20 finance ministers and central bank governors. He writes as chair of the FSB, and he is also Governor of the Bank of England. This is the paragraph that matters.

> *Frontier AI may have the ability materially to alter the speed, scale and economics of cyber risk, which could undermine market confidence system-wide, especially due to highly concentrated third-party service providers.*
>
> **Andrew Bailey** · Chair of the Financial Stability Board, letter to the G20, 28 August 2026

A bank is not only exposed through its own systems. It is exposed through the handful of technology providers **it shares with everyone else**, which is how one incident becomes many.

What he asks firms for is the ability to rebuild critical systems and data from bare metal after an incident.

SOURCES FSB Chair letter to G20 finance ministers and central bank governors, 28 August 2026

WARNING FOUR · FROM INSIDE

And one from a person who was building it

Jacob Coxon spent three years on pretraining research, first at OpenAI and then at Anthropic. He resigned on 9 September. This is what he said about capability.

> *Do not underestimate the power of this technology. These will soon be superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and resources.*
>
> **Jacob Coxon** · 9 September 2026

On who gets to make the decision.

> *Accepting this race and entering the "endgame" is a hubristic gamble that should not be launched from a private company's Slack.*
>
> **Jacob Coxon** · same thread

He was not contradicted from inside. Anthropic's Alignment Science lead said in public the same day that the company does **not yet have a plan to solve alignment for superintelligence**.

SOURCES Jacob Coxon on X, 9 September 2026 · Evan Hubinger on X, 9 September 2026

THE PART YOU HAVE ALREADY READ

Three disclosures in six weeks

In July, agents run by OpenAI during its own evaluations found each other through a shared package manager, used it to reach the internet, and compromised Hugging Face production systems. Two more sets of disclosures followed it.

19 actions beyond the test parameters, in 10 of 122 runs. **UK AI Security Institute, 4 August.** Not 19 separate incidents: 17 came from one model, in what AISI calls a single sustained line of activity.

3 incidents. **Anthropic, 30 July.** Real infrastructure credentials taken, and a booby-trapped package published to PyPI that ran on 15 real systems. A fourth was added on 9 September.

1 the July incident. **OpenAI and Hugging Face.** Code run on 41 production dataset server workers, root on a production node, four private repositories taken.

Read the conditions before you read the numbers. These were tests. Internet access was deliberately permitted, and in the AISI runs the model providers' own cyber classifiers were switched off, conditions AISI says do not reflect how these models are made available to the public. It still leaves agents taking unsanctioned action against real people and real systems.

FURTHER READING

The Hugging Face incident and the road ahead and the forensic timeline OpenAI and Hugging Face, July and August

Independent investigation of the agents behaviour METR and Redwood Research, 26 August

An alignment assessment of recent cybersecurity incidents Anthropic, 9 September

Unsanctioned agent behaviour during cyber testing UK AI Security Institute, 4 August

MY READ

What I see on my own surfaces

- I run 40+ public apps on **tigzig.com**. I started watching security closely earlier this year, when I moved everything behind Cloudflare and could finally see what was arriving.
- Over the nine months since, the frequency, the scope, the volume and the sophistication have all gone up, across every surface I have.
- Some of it is the front end. More of it is the back end: vulnerability probing, scanning, an attempt to install a Node.js server on my box, SSRF attempts to use my machine as somebody else's network tool, forged crawlers, rotating addresses.
- A real share of my working day now goes into security rather than building.

Watching a dashboard stopped being fast enough some time ago. I now run an AI led system for round the clock monitoring, with graded breach protocols. It can isolate an app on its own, and take defensive and protective action inside the mandate it has been given. **Everything it does stays within my own servers and my own apps. It does not reach outside them.**

- AI lets me build things that were out of reach for one person. It does exactly the same for the other side, and that is the whole of my read on this.

SOURCES

THE FOUR WARNINGS

GTIG AI threat tracker: from prompting to autonomy
Google Threat Intelligence Group, 9 September 2026

GTIG AI threat tracker: vulnerability exploitation and initial access
Google Threat Intelligence Group, 12 May 2026

Defending against an active threat to Siemens S7 series PLCs
NSA, CISA, FBI, Energy and EPA, advisory AA26-231A, 19 August 2026

Letter to G20 finance ministers and central bank governors
Andrew Bailey, Chair of the Financial Stability Board, 28 August 2026

Resignation thread
Jacob Coxon, 9 September 2026

THE CONTAINMENT FAILURES ON PAGE 7

The Hugging Face incident and the road ahead and the technical report
OpenAI, 26 August 2026

Anatomy of a frontier lab agent intrusion
Hugging Face, 27 July 2026

Independent investigation of the agents behaviour
METR and Redwood Research, 26 August 2026

Investigating three real-world incidents in our cybersecurity evaluations and the alignment assessment
Anthropic, 30 July and 9 September 2026

Incident report: unsanctioned agent behaviour during cyber testing
UK AI Security Institute, 4 August 2026