Claude

ANTHROPIC THREAT INTELLIGENCE REPORT

September 2026 · The cyber section

“Sophisticated attacks no longer require sophisticated attackers”

Eight months of activity, disrupted. A state espionage unit, a criminal crew and one person working on their own all ran the kind of campaign that used to need a team, and some break-ins were finished in two to three hours.

Key findings

+ What I am seeing in my own security logs

SOURCE

Detecting and countering misuse of AI: September 2026

Anthropic Threat Intelligence. These pages cover the cyber operations section only.

Amar Harolikar · Decision Sciences & Applied AI

You can no longer tell a state operation from one person working alone

A RUSSIAN ESPIONAGE GROUP

- → More than 20 organisations targeted
- → Ministries, embassies, defence firms
- → Hundreds of gigabytes taken
- → A campaign running 130 days

ONE FRENCH-SPEAKING OPERATOR

- → 42 organisations tracked as targets
- → Inside at least 14 of them
- → Tens of millions of rows in a doxxing tool they built, so people in one political movement could be looked up by name
- → A campaign running 36 days

> *“For threat intelligence investigators, sophistication has stopped being a reliable signal of who is behind an operation.”*

Anthropic’s point is that sophistication no longer tells the two apart, and that what still does is intent. Even just a year ago, they write, campaigns like this would have needed many skilled operators.

They left their AI rewriting their own malware until nothing could detect it

STEP 1

The malware goes out

Windows, Android and iOS tools, delivered through fake software updates and hijacked hotel wifi.

STEP 2

Agents watch for a detection

AI agents monitor whether any security product has flagged any part of the toolkit.

STEP 3

It rewrites itself

When something is caught, the agents modify and rebuild it on their own, with no operator involved.

STEP 4

It repeats until it is clean

The loop keeps going until the security products stop seeing it, and only then is the tool used again.

Anthropic’s reading is that this turns the cost around. A new detection **might once have slowed an attacker down**. Their own wording is that “at least in theory” it no longer will.

They built a machine to find new flaws, then went after government networks

THE MACHINE

- → Reads a security product's own code
- → Guesses where a weakness might be
- → Writes an attack to test the guess
- → Tries it in a private lab and loops until it works
- → Then attempts on those same products inside real organisations

WHY IT DOES NOT STOP

- → Target lists, stolen passwords and progress saved between sessions
- → A new session picks up mid-campaign knowing all of it
- → One lead agent splits the work across many others at once
- → Thirteen agents collecting open-source intelligence on a timer, nobody watching

One of those loops produced more than a dozen possible new flaws in a single month.

This was an espionage operation. The same operators targeted about fifty organisations and took student records, citizen records and access to a retailer's live systems. **Two of them were undergraduates at a Chinese university.**

One operator downloaded 1.8 million phone apps to read what was left inside them

COLLECT

1.8 million Android apps, pulled from several app stores

Run on a small fleet of rented cloud machines.

OPEN

Every app taken apart and scanned for passwords

Looking for keys a developer compiled in and forgot about.

SORT

Working keys arrive in a chat group in real time

Filed automatically into more than a hundred categories. A second pipeline did the same with stolen developer tokens.

This operator also left their own staging address and bot tokens exposed. That is part of how they were found.

Those two pipelines supplied **most of that operator’s break-ins**. Anthropic’s line: “everything connected to the internet is a potential target for exploitation.”

The same six steps, run again and again, and each one is something you can check

1 A key ships inside an app

One of hundreds of thousands of passwords compiled into ordinary mobile software.

2 It gets found and tested

Machine decompilation at volume, and the key is checked live to see what it opens.

3 The cloud account opens

A key that works gives access to the company’s wider cloud estate.

4 The build system is taken

Software pipelines backdoored, and every other stored password harvested.

5 Control sits inside production

The attacker’s own software now runs on the company’s live infrastructure.

6 Sale, drain and extortion

Keys sold on, payment flows intercepted, and the data used to demand a ransom.

Anthropic describe an operator who very often **may not directly understand** the target company, and is “deferring the specifics to the AI”.

Speed and scale are what changed

3 hrs

From one stolen developer login to full administrative control of a company's cloud

34 hrs

To take over 2,100 sets of login tokens spanning more than 40 corporate accounts

200

Customer organisations reached through the break-in at a single software supplier

> *"AI agents performed nearly all of the work."*

The report describes breaches finished in as little as two to three hours, and single operators handling dozens of victims at the same time.

Anthropic add two caveats. Humans still choose the targets and review what comes out, and **several of the worst compromises in the report were directed by a person at every step**.

Your AI keys are now worth stealing on their own

1 Loot

Stolen keys and accounts have a resale value in markets that already exist.

2 Compute

The attacker’s own work then runs at somebody else’s expense.

3 Cover

The activity is attributed to the person who legitimately owns the credential.

4 The discount that is not one

A group sold discounted Claude access. The traffic went to a different model and their software took the buyer’s details.

One actor attacked about thirty AI companies in four days, chasing an unreleased Claude model, and Anthropic say every path failed. In every instance the stolen keys came from customer environments, and their own systems were not compromised.

Their recommendation: treat AI keys and agent integrations with **the same level of seriousness as production credentials**, because attackers already do.

The ways in are the same ones as before

HOW THEY GOT IN

- → Stolen credentials
- → Unpatched edge devices
- → Exposed services
- → SQL injection
- → Phishing

WHAT IS DIFFERENT

- → The labour is handed to AI
- → It runs at machine speed
- → It runs in parallel
- → Dozens of victims at once
- → In one case the AI found new bugs itself

> *“None of the operations in this report depended on some entirely novel technique that defenders have never seen. Instead, the economics of the attacks have changed.”*

The research, the breaking in, the tool building and the data processing all used to separate a well-funded operation from everybody else. All of it is now delegated to AI.

Six questions, each one tied to something that happened in the report

1 Your own mobile apps

Is there a key or token inside anything you ship? That is where the biggest harvest started.

2 Your AI keys

Who holds them, where do they sit, and would you notice one being used from somewhere else?

3 What your suppliers can reach

One supplier’s access reached 200 of its customers. Ask what yours can reach in your systems.

4 Your edge devices

VPN gateways, firewalls and mail portals were the way in more than once. Check the versions.

5 Your code and build systems

Passwords in old commits and build pipelines were harvested repeatedly, and are easy to find.

6 Your first few hours

Some break-ins were finished in two to three hours. Ask how long you take to notice and act.

Anthropic’s framing: “The capabilities described in this report should be assumed to be **available to any actors who are motivated to use them.**”

FROM MY OWN LOGS

My own logs are moving the same way, and AI keys are hunted by name

I have run these tools for over two years, and for most of it I only had my own backend logs. Everything moved behind Cloudflare in January, and I could finally see what arrives at the edge. Frequency, scope, volume and sophistication have all gone up since.

WHAT CHANGED

→ A probe arrives as a slash and a company name, checking whether my box holds their data

→ One request per address across thousands of addresses, so a per-address limit does nothing

→ Badges get forged, so a request claims to be a search engine from a network it never uses

WHAT I DO NOW

→ An AI led watch, round the clock

→ Some sessions on watch duty, others on standby to investigate, connected to each other

→ It can isolate an app inside the mandate it has

→ It pages my phone while things are still happening

A technique gets refused and **a different one arrives within minutes** .. Two years of hardening, written up as a checklist on the next page.

MORE ON THIS

Two years of hardening my own tools, written up as a checklist

What is in it

132 items across 14 categories, each with the risk, a plain-English fix and the working code. It began as a few notes when I ran a handful of tools, and it grew every time something went wrong on a live service. tigzig.com/security

Four warnings on AI and cyber security

From an earlier post of mine. Warnings from Google's threat intelligence group, five US agencies including the NSA and the FBI, the chair of the Financial Stability Board, and a researcher who had just left a frontier lab. tigzig.com/post/four-warnings-ai-cyber-risk-sep2026

> The attacks that reach a bank are not the ones that reach me, and the sophistication is not comparable. **A lot of the doors are the same ones.** A key left in an app, an old commit, an unpatched gateway. I came to this as a data scientist rather than a security engineer, so I learned every one of these on a live service, the hard way, and I am still learning ..

THE REST OF THE REPORT

Cyber is one of seven areas, and these pages cover only that one

The full report runs to 154 pages and covers activity from December 2025 to August 2026. The other six areas in it are:

- **Influence operations**
- **Surveillance**
- **Conventional weapons**
- **Biological misuse**
- **Scams and fraud**
- **Illicit distillation**

In each case Anthropic say they disrupted the activity, banned the accounts and shared what they found with authorities and industry partners. They found no malicious activity on Claude Fable or Mythos.

SOURCE

Detecting and countering misuse of AI: September 2026

Anthropic Threat Intelligence, published 10 September 2026. Every quotation on these pages is from that report, and every number is theirs.