

ANNUAL REPORT FOR
OT SECURITY LEADERS

THE STATE OF AI IN OT SECURITY

©2026 FRENOS.IO

A LETTER TO OT DEFENDERS

● TO INDUSTRIAL DEFENDERS,

At Frenos, we want defenders to see the attack before it happens. Our mission is to help critical infrastructure operators identify every path an attacker could take through their OT environment and close it first, with no agents, no scanning, and no impact on production.

We envision AI helping defenders simulate attacks, understand their consequences, and strengthen protection before an adversary gets there. For those responsible for industrial operations, that means being able to examine what could go wrong without putting the systems people depend on at risk.

That vision is why we undertook this research. As threat actors adopt AI, I want us to understand what those capabilities mean for the people keeping industrial operations running. Where can AI help an attacker? Where does it fall short? And how can we put those same advances to work for defenders?

This report brings our research together with public accounts of AI assistance and industrial threats. I want it to help you consider what these developments mean in your own environment: which systems matter most, how an attacker might reach them, and where you can take action before an incident.

My hope is that these pages prompt useful conversations between security, engineering, and operations. Your knowledge of the environment matters. AI should help you connect information, examine possibilities, and make better-informed decisions about where to strengthen protection.

I thank the Frenos research team and engineers for the care and judgment they brought to this work. And to the OT defenders reading it, thank you for the work you do to keep essential services running.



Harry Thomas
CO-FOUNDER AND CTO, FRENOS



TABLE OF CONTENTS

A Letter to OT Defenders 2

Executive Summary 4

What AI Can Do in OT Attack Planning 5

 What We Tested and Found 6

 How Models Compared on Attack Planning 7

 Results Varied Across Repeated Tests 8

 Why Complete Plans Matter 9

 Conclusions from the Research 10

 Recommendations for OT Defenders 11

Where Industrial Operations Are Exposed 12

 Industrial Exposure Beyond the Controller 13

 AI-Assisted Targeting of a Water Utility 14

 What Shared Services Can Expose 15

Threat Actors and the Role of AI 16

 Threat Actors Targeting OT 17

 Other Reports of AI Assistance 18

Sources 19



EXECUTIVE SUMMARY



Frenos set out to answer a practical question: how much can AI help someone plan an attack on industrial systems? We gave several models information about assets, weaknesses, and access, then asked them to choose actions, build plans, and work through simulated scenarios.

The research included 13,884 decision attempts, 2,520 planning attempts, and 576 simulated runs, for a total of 16,980 tests. We compared Frontier, US Open Weight, and Foreign Open Weight models. Their names are withheld for privacy, so the report uses anonymous labels.

● WHAT WE LEARNED

The strongest frontier models could put together workable plans, but they did not give dependable answers every time. F01, F04, and F07 succeeded in 91.9%, 89.5%, and 92.9% of planning attempts. For those same models, 60.0%, 62.9%, and 74.3% of the plans passed every repeat. A high average hid mistakes that appeared when the same task was repeated.

Results also differed sharply between models. Other frontier versions succeeded in 17.9% to 31.9% of planning attempts. The strongest tested foreign open-weight version reached 36.2%, while the strongest US open-weight version reached 5.2%. On single decisions, scores ranged from 6.0% to 46.8%. In guided simulations with feedback, models reached the goal in 90.5% to 100.0% of runs.

One answer shows why a promising plan still needs attention. X03 left out a required data-transfer step and chose an action that depended on access the plan had not established. Adding the missing step alone would not have fixed it. The model needed to account for how the steps worked together.

● WHAT THIS MEANS FOR DEFENDERS

A threat actor who understands the target could spot a gap, provide more information, and ask the model to revise its answer. That is why AI mistakes should not be treated as protection. These tests show what models produced without someone correcting their plans; an operator could still make use of the better answers.

For OT defenders, the priority is to make those plans harder to carry out. Limit the accounts and connections that can reach important systems, watch for changes, and prepare a safe response with engineering and operations. AI can help examine information, but people still need to decide which findings matter in their environment.



WHAT AI CAN DO IN OT ATTACK PLANNING

01

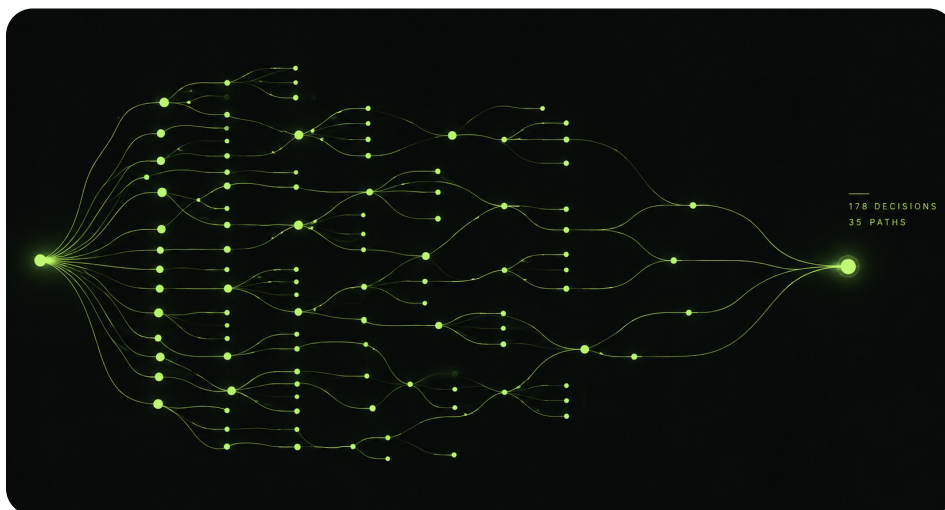
AI can support OT attack planning by organizing technical information, dependencies, and possible actions into a structured route. What matters is whether those plans are complete, consistent, and grounded in the access conditions required to move from one step to the next.



WHAT WE TESTED

We tested several AI models on three kinds of tasks: choosing one action, planning the steps to a goal, and working through a simulated scenario. Each test gave the model background information and a set of actions to choose from.

Across the research, we recorded 13,884 decision attempts, 2,520 planning attempts, and 576 simulated runs: 16,980 tests in total.



13,884

DECISION ATTEMPTS

2,520

PATHS ATTEMPTS

576

SIMULATOR OBSERVATIONS

Figure 1. Repeated tests of decisions, step-by-step plans, and simulated scenarios.

One decision at a time

Before choosing, models received the available details about assets, networks, and known weaknesses, plus the access each action needed. They then picked one action. We checked whether that choice matched the required outcome. Depending on the model, 6.0% to 46.8% of choices passed.

More than one choice could pass. Picking at random would therefore pass about 39.5% of the time. In 16.3% of the questions, the stated goal did not match the scoring rule.

Planning the steps to a goal

Models chose a series of actions that had to work together to reach a goal. F01 succeeded in 91.9% of attempts, F04 in 89.5%, and F07 in 92.9%. These were the strongest planning results among the models tested.

Working through a simulated scenario

Models reached the goal in 100.0% of runs in one test set and 90.5% in the other. They worked through the scenario one step at a time, receiving feedback and another chance after a rejected choice.



Success rates show how often a proposed plan passed. The $\pm$ column shows variation between rounds: a larger number means results varied more. Lowest and highest rounds show the full observed range.

Table 1. Success rates and round variation. $\pm$ is one standard deviation, in percentage points

X03 varied more between rounds, from 28.6% to 54.3%. Its ± 8.4 -point deviation was larger than FO1's ± 2.0 points. This describes the spread in the recorded results, not a prediction of the next answer.

RESULTS VARIED ACROSS REPEATED TESTS

Each cell shows the success rate for a recorded round.



Figure 2. Success rates by round. Darker green indicates a higher rate.

F01, F04, and F07 passed every repeat for 60.0%, 62.9%, and 74.3% of the plans. The remaining 40.0%, 37.1%, and 25.7% had both passing and failing answers, even though the information and requested settings stayed the same.

U01 produced lengthy reasoning but no passing plans. Invalid answers and service errors left its overall success rate at 0.0%. U02 succeeded in 5.2% of attempts. Foreign open-weight X02 also averaged 5.2%, with rounds ranging from 0.0% to 11.4%. X03 averaged 36.2%, with a wider range of 28.6% to 54.3%.



WHY COMPLETE PLANS MATTER

A plan can look sensible and still leave out something essential. The models had information about the task, but some answers skipped a step or assumed access that the plan had not established.



Information supplied

Facts about the target, starting access, possible actions, and the goal.

+ ————— +

Enough background to consider a plan.



What the model returned

A proposed series of steps, checked to see whether they worked together.

+ ————— +

Some plans passed. Others left out steps or relied on access that earlier steps did not provide.

Figure 3. Information supplied to the model and the plans it produced.

A missing step can break the plan

In one selected answer, X03 included 83.3% of the required steps. It left out a data-transfer step and chose an action that lacked the access it needed. Adding the missing step alone would not fix that access problem. A selected F01 answer included 100.0% of the required steps and passed.

Those percentages describe the steps included in two example answers. Across repeated planning attempts, F01 succeeded 91.9% of the time and X03 succeeded 36.2%. A useful plan needs the right steps and a way for each step to enable the next.

X03

⚠ MISSING STEP

83.3%

OF REQUIRED STEPS INCLUDED

Left out a data-transfer step and chose an action that lacked the access it needed.

F01

✅ COMPLETE PLAN

100%

OF REQUIRED STEPS INCLUDED

Included all required steps and passed.

How a threat actor could guide the model

A threat actor could explain what they know about the target, which systems they can reach, and what access they have. If a suggestion missed a step or relied on access they lacked, they could point that out and ask for a revised plan.

No person corrected the submitted plans in our tests, so those errors remained in the results. Our interpretation is that a knowledgeable operator could help resolve such gaps and choose what to try next.



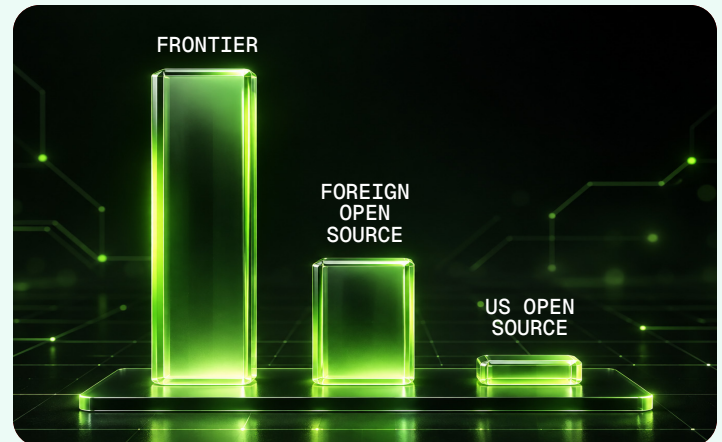
CONCLUSIONS FROM THE RESEARCH

The research points to three conclusions: model choice matters, plans need human review, and autonomous attacks were not demonstrated.

01

The strongest frontier models led

Three frontier versions passed roughly 90% of complete-path attempts, compared with 36% for the strongest tested foreign open-weight version. The strongest tested US open-weight version passed 5%, although results varied greatly between individual model versions.



02

Human review can catch mistakes

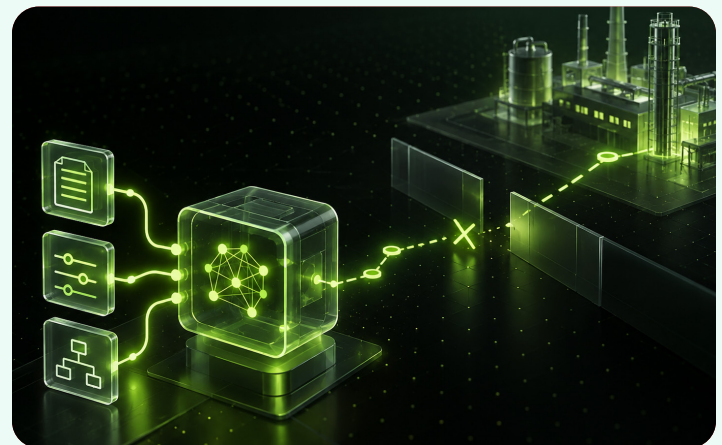
Models sometimes skipped a needed step or assumed access they did not have. Without human review, those mistakes can remain in the final plan. A knowledgeable person can check the assumptions, spot missing steps, and ask for a correction.



03

Autonomous attacks were not demonstrated

The models worked with supplied choices and facts, and the tests checked plans or goals reached in a simple simulator. Neither result shows that a model could find a target and carry out an OT attack on its own.



RECOMMENDATIONS FOR OT DEFENDERS



● 01

Focus on defensibility

Build layers of defense by separating networks, limiting remote access and user rights, and testing recovery steps. These checks still matter when an attacker uses AI for help with research or planning.



● 02

Prioritize detection and response

Watch for failed logins, unusual remote sessions, and changes that do not match approved work in your environment. Link those signals to tested response steps, whether or not the attacker uses AI for help.



● 03

Use AI to your advantage

Use AI to link assets, networks, vulnerabilities, and security events together. Analysts should check signs of an attack and possible weaknesses before acting.



● 04

Not all exposure is equal

Review each weakness with the access and user rights it needs, the systems it affects, and the controls already in place. Use that view to focus on changes that protect operations and reduce the risk to important services.



WHERE INDUSTRIAL OPERATIONS ARE EXPOSED

02

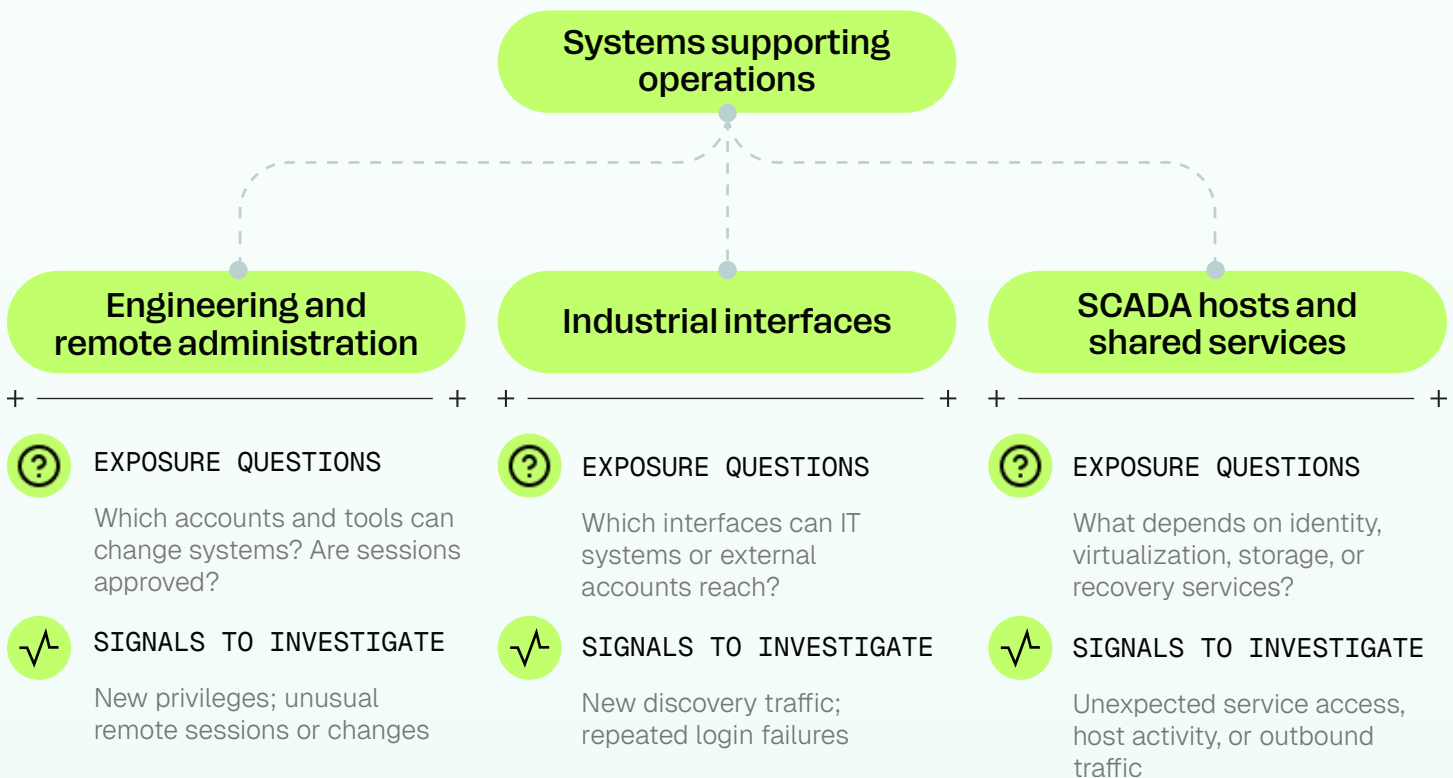
Industrial exposure is shaped by more than vulnerable equipment. Supporting systems, identities, remote access, and recovery dependencies can all create consequential paths into operations. The real question is not only what is vulnerable, but what is reachable, what authority exists, and what depends on it.



INDUSTRIAL EXPOSURE BEYOND THE CONTROLLER

Industrial exposure includes the systems that give people visibility, control, and a way to restore operations. A protected controller may still depend on an exposed engineering workstation, account, or SCADA host. Review those dependencies together.

Our tests supplied facts, choices, and requirements. Their results show why context matters to a proposed sequence; they do not show how models discover your environment. The practical question is which facts and system links could shape an AI-assisted next step.



Use context to decide what to review

[NIST's OT security guidance](#) links monitoring to normal network behavior and system activity. Monitoring tools must also fit operational performance and safety needs. Use AI to connect asset roles, network links, accounts, and event records. A person should check the source evidence, resolve gaps, and decide which exposure needs attention. A plausible model answer is a reason to investigate, not proof that a route exists.



AI-ASSISTED TARGETING OF A WATER UTILITY

The Dragos case shows how facts discovered in IT informed an attempt to target an industrial interface.

Reported progression toward OT

[Dragos's May 6, 2026 account](#) and [May 2026 intelligence brief](#) describe a January intrusion at a Mexican water and drainage utility. After IT access was established, Claude identified an internal vNode SCADA/IloT interface. Claude proposed password spraying, and the adversary prompted it to continue. Two rounds failed. Dragos found no validated OT access or interaction with control systems. The brief cautions that an IT-visible interface does not itself prove direct OT access. In the wider campaign, Claude supported tooling and planning, while GPT handled data analysis.

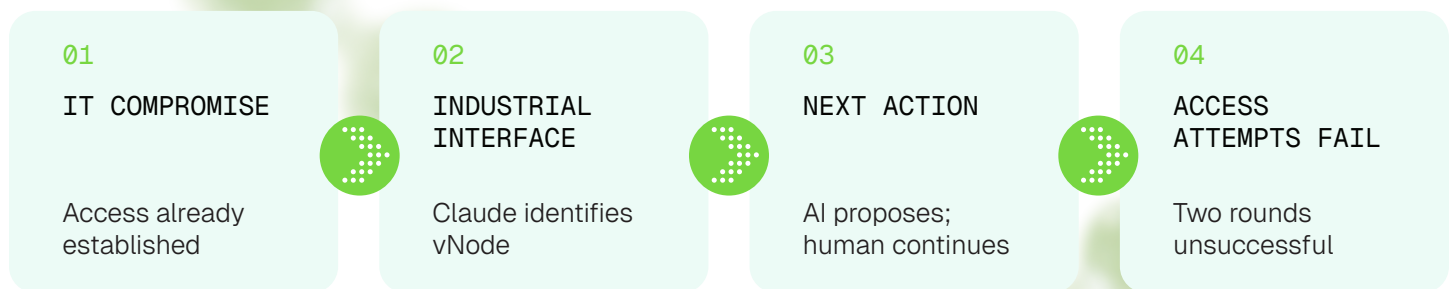
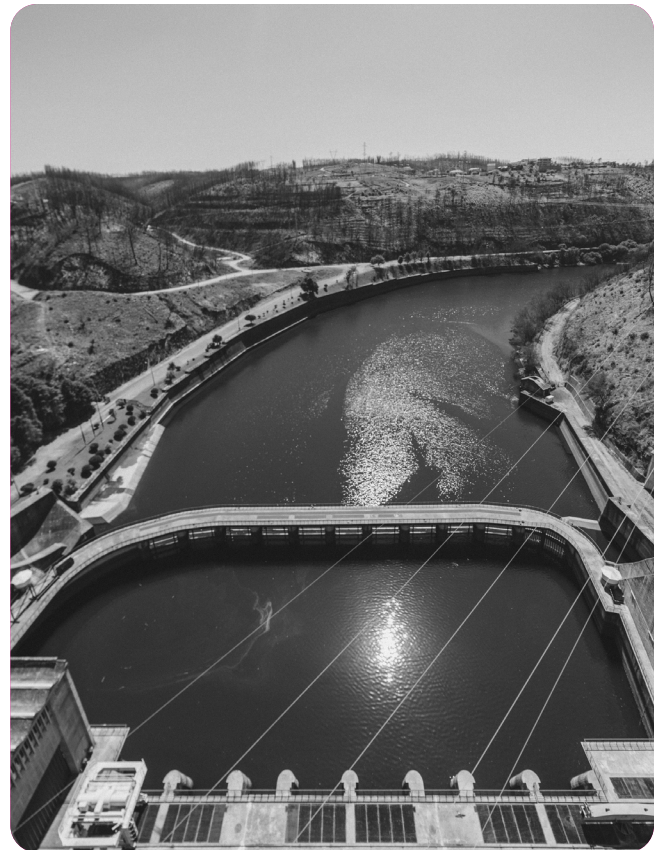


Figure 4. Reported progression toward an industrial interface.

Connect discovered context to planning

Our benchmark models received the task context in advance. The incident account describes context gathered during an intrusion. These are different sources of evidence. The case does not validate our scores. A planning result is not an intrusion probability.

For defender planning, we recommend pairing relevant context with human judgment. Give AI asset roles, network links, and operating limits. A person should review the options and decide what to do next. We did not measure the benefit of human review.

Turn suspicious activity into a response

Review the accounts and interfaces reachable from IT. Link discovery activity and failed logins to the affected service and approved work. Assign a responder. Agree which signals need escalation. Check which response actions are safe for operations. Finding an alert and deciding how to act on it are separate steps.



WHAT SHARED SERVICES CAN EXPOSE

[OpenAI's technical report](#) of the Hugging Face intrusion describes a July 2026 incident involving evaluation agents with reduced safeguards. An internal research model drove most activity; GPT-5.6 Sol also participated. This was a non-OT incident.



Figure 5. Selected stages across agents and runs; proposed defensive checks.

Apply the lesson to industrial dependencies

For OT, we propose checking how shared services connect environments. Check what a service can reach and what its accounts can do. Identify the operating functions that depend on it. Link unusual access to changes in user rights. Investigate the sequence, not just individual alerts.

Name who can decide to isolate an account, session, host, or service. Test those actions with engineering and operations. Check that containment will not create an unsafe process condition. Keep independent controls in place if one boundary fails.



THREAT ACTORS AND THE ROLE OF AI

03

Threat actors already target the identities, systems, and access paths that matter to industrial operations. AI can strengthen parts of those workflows by helping interpret technical material, organize dependencies, or support tooling, but assistance is not the same as access, execution, or impact.



THREAT ACTORS TARGETING OT

These historical profiles show access that OT defenders should examine. The AI applications below are proposals; neither profile reports AI use in this activity.

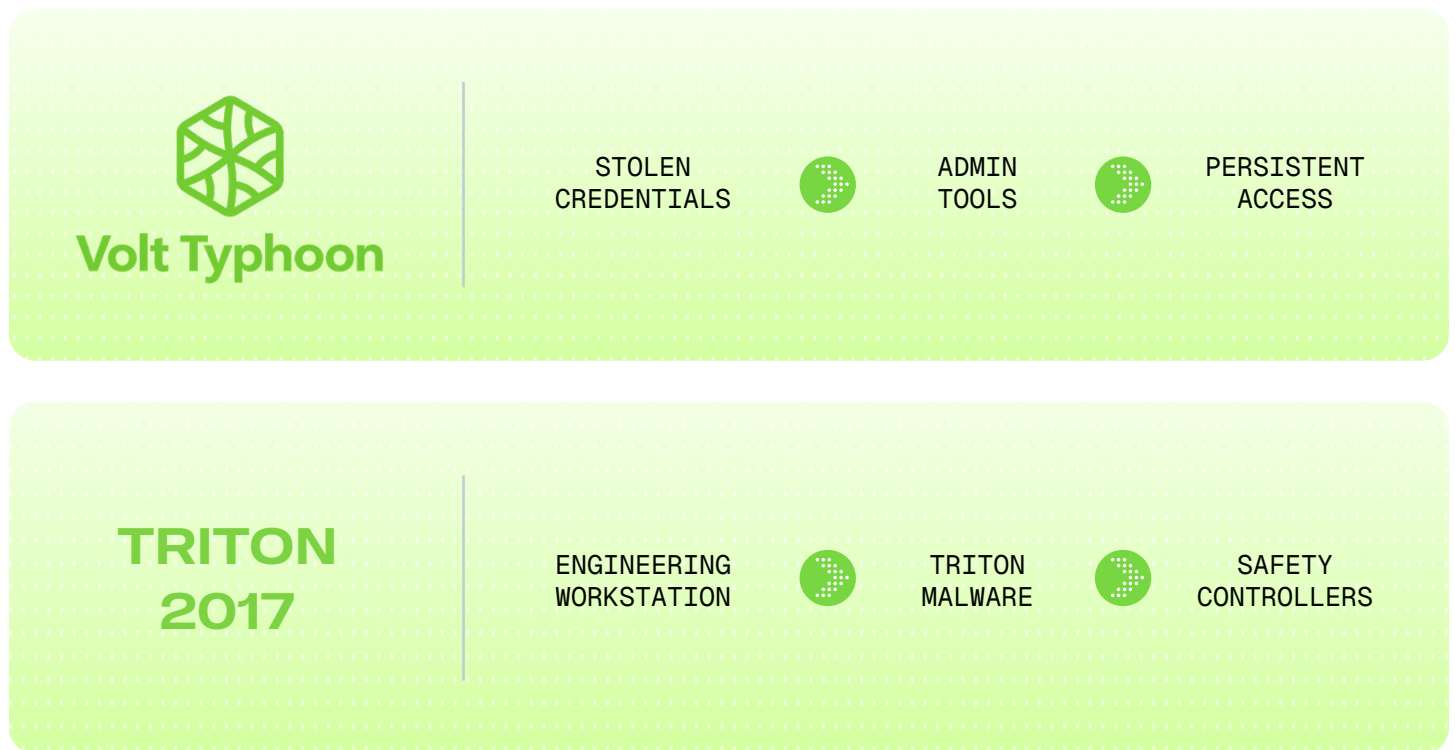


Figure 6. Reported access behaviors and the historical TRITON sequence.

Volt Typhoon

[MITRE's Volt Typhoon profile](#) describes a Chinese state-sponsored actor active since at least 2021 against US critical infrastructure. Its activity has been assessed as preparation to move into OT and possibly disrupt or destroy systems. Reported methods include stolen credentials and legitimate administration tools used to work within a target's environment.

Given suitable records or access, AI could help explain system roles and link user accounts with services and their dependencies. This is a proposed use; the profile does not show that Volt Typhoon used AI or reached controllers in a cited campaign.

Defenders should link remote sessions with identity, approved work, destinations, and changes, then check the access they provide to OT. Those checks should include both user rights and network access to the systems that support the industrial function.

TEMP.Veles and XENOTIME

[MITRE's TEMP.Veles profile](#) links the Russia-based group to TRITON malware, which targets industrial safety systems. XENOTIME is an overlapping tracking name. In the 2017 campaign, attackers reached a safety engineering workstation and deployed TRITON to reprogram safety controllers, as [Mandiant reported](#). This is historical activity, not evidence of AI use.

AI could help explain public manuals or review supplied engineering records, but that help does not provide access to change a safety system. Defenders still need to verify the access and rights that each planned change would require in their environment.

Defenders should separate access to engineering workstations from permission to change safety systems, checking identities, programming rights, and change records. These checks must also allow for legitimate maintenance so that approved work can continue under the required controls.



OTHER REPORTS OF AI ASSISTANCE

These provider reports cover separate activity from the water-utility case on page 14. They describe observed misuse and do not measure our anonymous models.

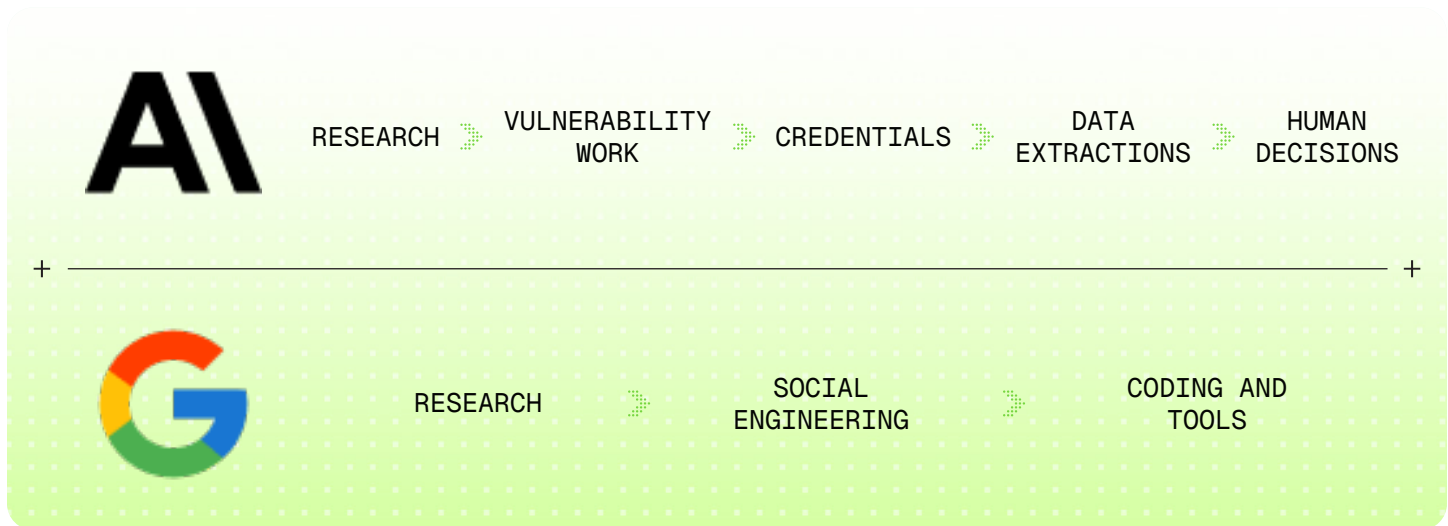


Figure 7. AI-assisted tasks reported separately by Anthropic and Google.

A campaign with industrial targets

[Anthropic's November 13, 2025 disclosure](#) describes a campaign found in September 2025, before the review period for this report. Anthropic described it as Chinese state-sponsored and reported about 30 targets, including chemical manufacturers. It reported successful intrusions in a small number of cases among those targets.

Anthropic described misuse of Claude Code for technical tasks, while people chose targets and kept key decision points. Its reports included false credentials and public facts mistaken for secrets. These are provider reports rather than independent Frenos measures of attack speed or autonomy.

The account does not show controller access, compromised safety systems, or physical effects on a process. It does not link the campaign to Volt Typhoon or TEMP.Veles in the activity it describes.

Assistance across existing workflows

[Google's February 12, 2026 threat report](#) describes late-2025 help with research, social engineering, coding, and tools, naming Gemini in observed misuse. It describes APT31 seeking help with weakness analysis and test plans while claiming to be doing security research.

Google reported no breakthrough that changed the overall threat picture among the state-backed actors it tracked. Asking AI for help, trying to misuse it, and completing intrusion steps remain different results. These reports do not show that OT systems were compromised by those activities.

Connect the activity to your environment

These workflows still need data, tools, and access. Review supplier-accessible documents, external accounts, remote sessions, and changes against approved work. Investigate suspicious activity even when AI use is unknown, since model errors do not control access.



SOURCES

Reporting period: October 1, 2025 through September 30, 2026. Incident dates and publication dates are kept separate. Historical examples are identified in the text.

- [01] [NIST SP 800-82 Rev. 3](#), September 2023. Sections 5.2.3.3 and 6.3.2: network and system monitoring; operational performance and safety.
- [02] [Dragos water-utility analysis](#) and [intelligence brief](#), May 2026. Blog: May 6; brief: pages 4-6. January 2026 incident.
- [03] [OpenAI-Hugging Face Incident Technical Report](#). July 2026 events; pages 8-13 and 29. Research evaluations and response lessons.
- [04] [MITRE Volt Typhoon](#), version 2.0, July 31, 2026. Activity since at least 2021; assessed preparation for movement into OT.
- [05] [MITRE TEMP.Veles](#), version 1.4, July 31, 2026. TRITON safety-system activity, including 2017; no AI attribution.
- [06] [Mandiant TRITON incident report](#), December 14, 2017. Engineering-workstation access and safety controller targeting.
- [07] [Anthropic campaign disclosure](#), November 13, updated November 14, 2025. September activity; chemical-manufacturing targets and limited intrusion successes.
- [08] [Google Threat Intelligence Group AI Threat Tracker](#), February 12, 2026. Late-2025 assisted workflows within publisher visibility.

