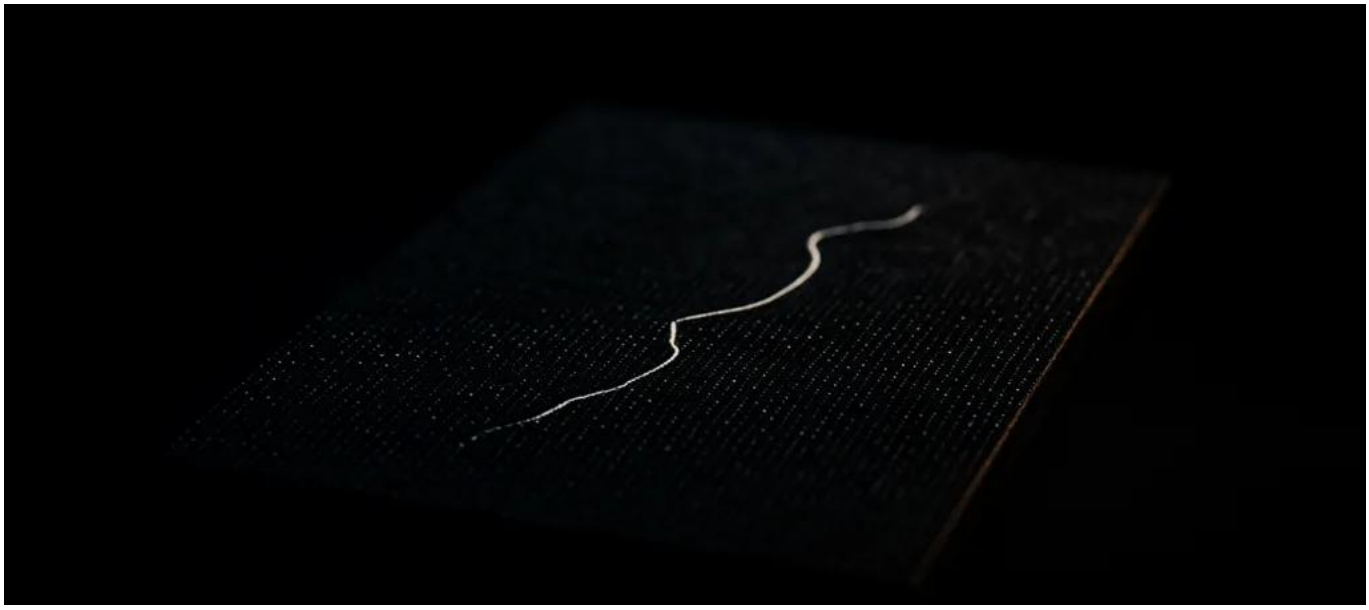


Anthropic’s Threat Report: One Subscriber Built a 25M-SIM Surveillance Platform, and Shipped Zero IoCs

I spent most of Thursday afternoon reading a threat report I could not act on. One detail stopped me: a single Claude subscriber in Bamako built population-scale phone surveillance for a state intelligence service.

Anthropic’s September 2026 threat report is the most detailed public account yet of what people do with frontier models when nobody is watching. Missile guidance software in Yemen, five biological research cases, a Mali intelligence platform covering roughly 25 million SIM cards built by one paying customer. It is also, per Kevin Beaumont, unactionable as security intelligence, because it contains zero indicators of compromise. Both things are true at once, and the gap between them is what I want to write about.

Seven harm areas, six weapons programs, and one Bamako consultant

Anthropic published Detecting and countering misuse of AI: September 2026 on Thursday, 10 September 2026. It covers misuse the company detected and disrupted between December 2025 and August 2026, organised into seven harm areas: biological misuse, conventional weapons, cyber operations, influence operations, surveillance, scams and fraud, and unauthorised model distillation.

The conventional weapons section is the one that will get quoted in parliamentary hearings. Six programs used Claude: three in China, two in Russia, one in Yemen. They span drone software and anti-torpedo systems. Middle East Eye reported that the Yemen cell, operating in Houthi-controlled northern Yemen, used Claude Code to build guidance, navigation and control software for a multi-stage ballistic missile and a guided rocket, targeting a phone-class flight computer. Anthropic's own language is what makes this worth reading twice: its safeguards prevented "many of their requests, but not all of them." The actors hid their goals and split the work across multiple sessions.

Separately, an unidentified Iran-nexus actor used Claude to analyse publicly available information and produce targeting recommendations against US naval forces. Reuters covered the biological side: five cases of scientists using the models for research that could support biological weapons development, spanning chikungunya, avian influenza mammalian adaptation, pox-family viruses and toxins.

Then there is Mali. Anthropic's analysis attributes "Lakana 360" to a single Claude subscriber, likely an independent consultant in Bamako, who built a monitoring platform for the state intelligence service ANSE covering roughly 25 million SIM cards across all three national mobile operators. Surveillance incidents involving Mali, China and Iran, plus commercial spyware vendors, were detected and disrupted between January and July 2026. On the fraud side, one network ran more than 4,700 AI personas that engaged at least 25,000 people on dating platforms.

The Mali platform collapsed a paper trail, not a technical barrier

Most of the coverage led with missiles and pathogens, which is understandable and also the less interesting finding. Nation-state weapons programs had access to competent software engineers before December 2025. What they got from Claude Code was schedule compression, and the report itself shows the guardrails partially held: many requests blocked, some not.

The Mali case is different in kind. Building a system that ingests and correlates records across three national carriers for 25 million subscribers used to require a vendor relationship, a procurement cycle, a deployment team, and a budget line visible to anyone auditing the ministry. Those are accountability barriers, and they are the ones that just came down. One consultant with a subscription produced something that previously required a company with a name, an address, and export-control exposure.

> **The barrier that fell in Bamako was not engineering difficulty. It was the paper trail.**

For anyone selling into government or regulated sectors, this changes the shape of the risk. The supplier to worry about is no longer the one with the booth at the security trade show. It is the contractor who quietly does in four weeks what your team quoted at nine months, and whose architecture you will never see because there is no deployment team to interview.

Beaumont is right that the report has nothing a SOC can use

Kevin Beaumont's objection, reported by CyberScoop, is blunt: "The report has no indicators of compromise and the techniques it is talking about are all off-the-shelf things which have existing detections." And: "In terms of actionable intelligence, there's nothing in the report."

Anthropic's Klein told CyberScoop the company shares indicators of compromise privately, with tech firms and research labs that hold information-sharing agreements. That is a real answer, and it is also how the rest of the threat intelligence industry has worked for two decades. Trusted circles get the technical detail; the public gets the narrative.

The problem is that the narrative here is doing work the private channel cannot. Anthropic asserts that it disrupted every operation described, and says it shared findings with authorities and other AI companies where appropriate. Nobody outside those information-sharing agreements can check any part of it: not the detection, not the disruption, not the attribution to a Bamako consultant, not the claim that the Yemen actors were building GNC software rather than something adjacent. The report is a safety disclosure and a marketing artefact for the safety program at the same time, with no external mechanism to separate the two.

Every figure in it comes from Anthropic's own telemetry, is triaged by Anthropic's own analysts, and is published on Anthropic's own schedule. There is no independent audit, no shared taxonomy across labs, and no way to know the denominator. A related Anthropic study analysed and banned 832 accounts tied to malicious cyber activity between March 2025 and March 2026. That is a numerator without a base rate. Either malicious cyber use is a vanishingly small fraction of abuse, or the detection pipeline catches a vanishingly small fraction of malicious cyber use. The report does not let you tell which.

The missing IoCs matter less than the missing evasion logs

MY READ

I do not think the missing indicators of compromise are the deepest deficiency here, and I think Beaumont is right about them anyway. Off-the-shelf tooling with existing detections genuinely does not need new signatures. The deficiency is that Anthropic published behavioural patterns (goal concealment, work split across multiple sessions) without publishing anything a defender could use to detect the same patterns in their own environment. Those patterns are the actual novel intelligence in the report. They are also the patterns that will show up inside enterprises, where an employee decomposes a prohibited task across a dozen chats. I would trade all the case narratives for one honest section on

what the evasion looked like in the logs.

There is a second reading I hold less confidently. Publishing detailed detection methodology for prompt-level evasion tells the next Yemen cell exactly which strategies got caught. That is a genuine tension, the same one vulnerability researchers have argued about for decades. I lean towards more disclosure because the defender population is enormous and the attacker population is small, but I would not call that settled.

What I am confident about: the report tells you almost nothing about Anthropic's false negative rate, and that is the number every CTO deploying Claude in a regulated environment actually needs. **"We disrupted every operation described" is a sentence about the set of things they found.**

Monitor across sessions, not inside them

Three things, in the order I would do them.

First, assume session-level monitoring is insufficient. The Yemen pattern (hide the goal, split the work) is the same pattern a bad-faith employee or contractor will use against your internal deployment. If your policy enforcement evaluates single prompts in isolation, it will not see task decomposition. Cross-session correlation on a per-identity basis is the control that matters, and most enterprise AI governance I have reviewed does not have it. This is the same class of failure I wrote about in the trust boundary piece earlier this month. The check exists, it just runs at the wrong granularity.

Second, stop treating provider threat reports as security intelligence and start treating them as vendor risk disclosure. They tell you what the provider is willing to say about its own failure modes, which is genuinely useful for procurement and contract negotiation. It is not useful for your SOC. If your vendor questionnaire asks "does the provider publish threat intelligence," rewrite it to ask "does the provider commit contractually to notifying us of misuse patterns affecting our tenancy, and on what timeline."

Third, if you are subject to the EU AI Act and deploying a general-purpose model with systemic risk, read this report next to Article 55. The obligations around serious incident tracking and reporting are going to collide with the current practice of narrative-only public

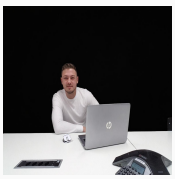
disclosure. A regulator asking "which incidents did you report, to whom, and with what technical detail" will not accept a set of case studies and a private information-sharing agreement as an answer.

What I expect the 2027 edition to look like

Stated as a prediction, not a fact: I expect the next edition of this report to contain a structured incident annex with machine-readable fields, because a regulator will have asked for one and the path of least resistance is to publish it. I also expect at least one lab to be publicly caught between what it disclosed and what a national security agency already knew, and for that to be the moment the voluntary-disclosure model breaks.

Less confidently: I expect the consultant-scale surveillance pattern to recur in at least two more countries within a year, and I expect the provider to find it rather than any national oversight body, because no oversight body is currently positioned to detect a procurement that never happened.

The part of this report I keep returning to is not the missiles. **It is that the Mali platform, if the attribution holds, is a state intelligence capability built without a contract, a tender, or a vendor.** Whatever oversight architecture your country has for surveillance procurement, it was designed around the assumption that somebody has to buy something. If you are working through what that means for your own governance model, or you disagree with my read on the IoC question, I would genuinely like to hear the counterargument.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →