

Analytical article

Code Does Not Fly on Its Own

An analytical paper based on open sources that separates AI use, produced outputs, independent verification, and operational capability, placing the northern Yemen case in regional history without exceeding the limits of attribution or evidence.

Article date

September 22, 2026

Author

Safwan AlModhayan**Keywords:** AI impact measurement · cybersecurity · multi-agent systems · Yemen · threat attribution · AI governance

English edition adapted from the author's Arabic open-source analysis, with evidence through 22 September 2026. It is not an independent digital-forensic investigation, a new experiment in weapons performance, or peer-reviewed research.

Abstract

This paper examines the claim that artificial intelligence increases technical capability in security-sensitive settings, beginning with Anthropic's report on a cluster in northern Yemen that used Claude Code in software work connected to weapons development. It proposes separating four stages that public accounts often collapse: observed use, produced output, independent verification, and accepted outcome. The evidence establishes software activity, a test that appeared to fail, and a return to the service for failure analysis. It does not establish a confirmed Houthi identity, deployment of an operational system, or the magnitude of any causal uplift from AI. ^[1] ^[2]

The paper places the case within a regional record of cyberattacks, missile and drone capabilities, and effects on energy and shipping, while keeping each event within the boundaries of its source. It then considers productivity measurement, the verification tax, the limits of multi-agent review, and the model provider's field of view. It derives institutional requirements relevant to Saudi Arabia: least privilege, independent testing, a decision record, and measurement of total cost per accepted outcome. These are analytical findings from an open-source narrative review, not results from experiments conducted by the author.

Method and Evidence Boundaries

The paper uses an analytical narrative review of open sources. Primary sources, or sources closest to the event, were preferred where available: company reports about activity inside their services, UN and government documents, statements by affected organizations, and research papers. Reputable journalism was used for party responses and developments without a complete accessible primary record. The evidence cutoff is 22 September 2026. Information that became available later does not enter the conclusions, even if it was known by the time of publication.

This review does not claim systematic coverage. It did not run a database-search protocol or a quantitative study-quality assessment. It has no access to Anthropic's raw logs, the cluster's files, independent test data, or field evidence. Provider statements are therefore treated as direct evidence of what the provider saw inside its platform, not as independent verification of performance beyond it. Statements by governments and parties to a conflict are treated as attributed accounts to be weighed, not as automatically neutral facts.

The analysis used a claim ledger: each claim was assigned a source, evidence type, visibility limit, and attribution status. Details that could become practical guidance for developing or targeting a weapon were excluded, and technical gaps were not filled by speculation. Where accounts conflict, the text preserves the conflict instead of settling it with language stronger than the underlying evidence.

Historical cases were selected for analytical function: to distinguish disruption from espionage, allegation from conviction, physical effect from impression, and support from direct command. The list is not a census of regional incidents. The final wording was

checked against the limits of each source so that an older event would not be used to compensate for missing evidence in the new case.

1. The Original Case, Before the Metaphors Begin

In its September 2026 report, Anthropic described a cluster in northern Yemen that used Claude Code in parts of software work associated with several weapons-development programs. The company said the users divided writing, research, and review among several instances, conducted a launch test that appeared to fail, and then returned to the service to analyze what had happened. It also said it banned the accounts and had evidence that the users had created an offline simulation tool. ^[1]

The report's most consequential limit is that Anthropic had no evidence the actors succeeded in fielding an operational device. That statement does not make the activity unimportant: according to the source, it reached a physical test and an attempt to learn from failure. It does prevent the observation from becoming a claim about a ready weapon or improved accuracy. The open record does not establish that the software output was correct, what the actors would have achieved without the tool, or whether the offline tool remained useful after the accounts were closed. ^[1]

Anthropic did not call the users Houthis. It located the cluster in northern Yemen. The Associated Press presented the Houthi connection as an inference from geography and context and carried a response from Hazam al-Assad, a member of the group's political bureau. The response is necessary evidence about a contested narrative; it neither independently verifies nor disproves the technical data. The methodological conclusion is exact: location is stated in the provider report, organizational identity remains unconfirmed, and operational success is not established. ^[2]

2. Regional History Did Not Begin in 2026

The Houthi movement emerged from local conflicts in Yemen before taking control of Sana'a in 2014. The Saudi-led coalition intervened in 2015 in support of Yemen's internationally recognized government. That history cannot be reduced to a new technical relationship, but the external factor cannot be edited out either. Reports by UN

expert panels present evidence and assessments that external materiel, technical assistance, and training contributed to the development of Houthi capabilities, while Iran has continued to deny arming the group. ^[12] ^[16]

Support, knowledge transfer, coordination, and direct operational command are different claims. A source may establish components or technical similarity without establishing where a decision for a particular operation was made. It may establish training or technical assistance without proving that the supporting party directed every later use. Causal analysis fails when ‘Iran and the Houthis’ becomes a single variable with no period, task, or type of relationship.

The record of launches and attacks against Saudi Arabia and in the Red Sea establishes that this debate concerns material capabilities that existed before the new case. The United Nations documented large numbers of missiles and drones and attacks on civilian and economic facilities; later Red Sea operations widened the effects to international shipping. None of those records establishes that Anthropic’s cluster was a direct organizational continuation of any earlier unit. History makes the hypothesis important. It does not supply evidence the new source did not publish. ^[9] ^[10] ^[11]

3. From Disk Wiping to Knowledge Theft

Saudi Aramco suffered the Shamoon attack in August 2012, disrupting tens of thousands of systems, and updated waves targeted Saudi organizations in 2016. The sources distinguish findings announced by Saudi investigations from attribution by intelligence assessments or security vendors to actors linked to Iran. The substantial disruptive effect is established; the degree of attribution must remain attached to its source rather than being presented as a comprehensive judicial finding. ^[18]

In 2017, Mandiant assessed that APT33 worked at the behest of the Iranian government and documented interest in aviation, energy, and petrochemical sectors, including organizations connected to Saudi Arabia. In 2022, participating US and UK government agencies described MuddyWater as a subordinate element within Iran’s Ministry of Intelligence and Security. The cases differ in source, facts, and confidence. A similar political background does not turn them into one incident. ^[19] ^[20]

On 18 August 2026, the US Department of Justice announced a superseding indictment charging seventeen members of the Mabna Institute, nine of whom had previously been charged in 2018. The indictment alleges a broad campaign to steal academic and intellectual-property data for Iranian entities. Its contents remain allegations unless adjudicated. The case nonetheless illustrates a strategic point: research knowledge is itself a target. ^[7]

This record explains why a model provider now belongs in the threat-intelligence landscape. The traditional questions were: who entered the network, what did they steal, and what did they disrupt? A new question joins them: what capability did a user try to build with the platform's assistance? Similarity between the questions does not license us to merge actors or cases without evidence. Attribution links a specific event to a specific effect through a specific source; it is not the selection of whichever name is most available in political memory.

4. When Risk Leaves the Screen

On 14 September 2019, Saudi Aramco said the attacks on Abqaiq and Khurais suspended production of 5.7 million barrels of crude oil per day. The attack and its operational effect are established in the affected company's statement. Responsibility, equipment origin, launch location, and political decision are separate questions requiring separate evidence. ^[21]

On 17 January 2022, a Houthi-claimed attack in Abu Dhabi killed three civilians. The economic reach of such insecurity later widened through attacks in the Red Sea. The International Monetary Fund estimated that Suez Canal trade volume fell by about half in the first two months of 2024 compared with the same period in 2023, as shipping was rerouted and delivery times lengthened. The figure describes a defined period and is not a current indicator for 2026. It does show how a security event can move from the point of attack into insurance, inventory, and supply chains. ^{[22] [23]}

In September 2026, the World Food Programme expanded its response in Yemen amid renewed fighting and displacement. Neither WFP nor Reuters connected that escalation to the Claude case, and editorial analysis must not do so. The humanitarian context is included to show that capability cannot be measured only through a technical component's performance. Social effects may appear in food access, movement, and continuity of services. ^{[4] [5]}

From a systems perspective, capability is a chain of dependencies: information, software, hardware, integration, testing, operations, and logistics. Public analysis does not need to describe targeting techniques or sensor combinations. It is enough to recognize that weakness at any link can constrain the outcome, and that improvement in software work does not demonstrate improvement across the full chain.

5. Define Uplift Before Measuring It

The claim that AI ‘increased capability’ is incomplete until the outcome has been specified. Uplift might mean shorter task time, fewer expert hours, higher quality, more attempts, or a larger set of feasible projects. Those measures are not synonyms. Output volume may rise while reliability falls. Writing may shrink while cost moves into review. One technical component may improve while testing and acceptance continue to constrain the whole project.

In causal terms, the quantity of interest is the difference between the outcome with the tool and the outcome without it. History gives us one realized path; it does not run the same event twice. Observing a user ask for assistance does not tell us how long the task would have taken otherwise. Observing a failed test does not reveal whether the tool moved the actors closer to success or farther away. The northern Yemen case therefore supplies no numerical coefficient for capability uplift.

A narrower hypothesis can still be stated. Models may reduce friction in knowledge work, making search, drafting, comparison, and iteration faster or cheaper. If that is confirmed, the strategic effect may be more attempts rather than guaranteed success in each attempt. This is a testable hypothesis, not a result that can be extracted from one provider report.

On 9 September 2026, Google’s Threat Intelligence Group described some adversaries moving from basic prompting to agentic workflows and connected automation. The report supports a wider trend toward organizing work through agents. It does not independently verify the Yemen case or establish that the arrangement outperforms human work or a single tool. ^[3]

6. When the Stopwatch Disagrees with the User

METR offers a useful example of the distance between perception and measurement. In a randomized study released in July 2025, sixteen experienced developers completed 246 tasks in repositories they knew. Within that study, they took 19 percent longer when AI tools were available, even though they believed after the experiment that the tools had made them faster. The finding does not show that AI slows every developer. It applies to particular users, repositories, tools, tasks, and a particular period. ^[26]

In February 2026, METR published a design update explaining that wider agent adoption had changed which developers would participate and which tasks they would submit, weakening the earlier design's ability to estimate later uplift. The researchers believed late-2025 tools likely accelerated some work, but called the evidence on magnitude weak because of selection effects and difficulties measuring parallel agent work. Serious inquiry gives neither an old result permanent residency nor an inconclusive update the status of victory for the opposite story. ^[27]

A later conceptual analysis distinguished uplift on old tasks, uplift on new tasks people choose after the technology becomes available, and uplift in total value. A tool may change the list of projects worth attempting rather than only the speed of the old list. That distinction is especially relevant in security: the tool might make previously uneconomic attempts worth trying without improving the quality of every result. ^[28]

The lesson is methodological. A survey measures belief; a time log measures a defined duration; a quality test measures properties of an output; an operational outcome measures value in a real environment. None of those measures should be allowed to impersonate all the others.

7. The Verification Tax and a Reality That Does Not Read the Deck

Take a civilian illustration: a project needs two hours to produce a software draft and eight hours to review, test, and integrate it. Even if the first two hours vanish, the project's duration does not. This is an illustrative calculation, not an experimental finding. It exposes how easy it is to celebrate acceleration in the visible stage while the actual constraint remains in testing, approval, and operation.

Sensitive work poses two questions. Was the system built according to the specification? And did the specification represent the real problem? Software can pass a correctly executed test built on incomplete assumptions. A simulation can be internally consistent while the physical world breaks an assumption that never entered it. ‘The simulation passed’ therefore means little without its scope and exclusions.

Richard Feynman ended his appendix to the Challenger report with a line that remains exact: ‘For a successful technology, reality must take precedence over public relations, for nature cannot be fooled.’ The line matters for the test it demands. Where is the test capable of contradicting our story? If a system produces the answer, explains why it is correct, and then grants itself a certificate of success, it has created a linguistically coherent circle, not independent verification. ^[29]

The evidence ladder therefore has four steps: observed use, produced output, independent verification, and accepted outcome. Moving from one step to the next requires new evidence. In the provider account, the northern Yemen case reaches use and outputs, followed by a test that appeared to fail. It does not reach demonstrated operational acceptance.

8. More Agents Do Not Guarantee More Independent Errors

Dividing work among one agent that writes, another that researches, and a third that reviews can be useful. Independence does not follow from the number of windows. Agents that share the same data, success criterion, or false premise may simply agree on the premise. A reviewer that sees the answer before constructing a test may end up improving its presentation rather than challenging its result.

The arXiv preprint ‘Why Do Multi-Agent LLM Systems Fail?’ analyzed 1,642 execution traces across seven frameworks and identified fourteen failure modes in system design, inter-agent misalignment, and task verification and termination. Experts developed the taxonomy from an initial sample; an automated annotator, validated against human labels, extended labeling to the full dataset. Experts did not manually label all 1,642 traces. The study does not assess the Yemen cluster. It does reject the shortcut from ‘several agents’ to ‘reliable review.’ ^[25]

Real review requires differences in evidence and task. The reviewer begins from a clear specification, can reach a primary source or test not produced by the writer, and has authority to say that the evidence is insufficient. Its value can be measured by defects found before release, not by the number of comments added to a conversation.

Reliability engineering asks more than whether the system completed a task. Did it detect the condition under which it should stop? Did it preserve a trace of the decision? Did it communicate uncertainty clearly? Independence is a property of design and testing, not a role name in a workflow diagram.

9. Outputs Remain; the Provider Sees Its Window

Anthropic said the users had created an offline simulation tool before their accounts were banned. The tool's existence does not prove that it was correct or useful. It does mean that ending access does not automatically retrieve outputs. An institution that closes an external subscription may retain files, code, or decisions produced with its help. An intervention should therefore be evaluated through three questions: what access stopped, what outputs are known to have left, and what evidence exists of later use? ^[1]

At the same time, model-provider logs offer visibility unavailable to many other security tools. A provider may see parts of the problem-solving sequence, changes in requests, and tool use. That justifies treating provider reports as important sources. It does not grant knowledge of what happened after a file was downloaded or inside an environment that never passed through the service.

Anthropic also says the selected cases are not representative of normal use; they were chosen because they were notable or novel. Published case counts therefore cannot estimate prevalence. More disclosures may reflect more activity, better detection, a changed publication policy, or a mixture of all three. Without a denominator, a count cannot become a rate. ^[1]

Journalism does not solve the problem through multiplication alone. Fifty articles repeating the same disclosure may extend the explanation and add interviews, but they do not become fifty independent witnesses to the original claim. Evidence independence must be assessed claim by claim, not logo by logo.

10. Saudi Governance That Measures Full Value

Saudi Arabia has direct reason to care about misuse of AI and equally direct reason to use AI to improve productivity and resilience. Governance should not become a synonym for prohibition. The requirement is a form of use that can demonstrate value, contain error, and sustain service when a tool fails or changes.

On 5 July 2026, the National Cybersecurity Authority opened a public consultation on AI cybersecurity guidelines covering governance, defense, resilience, and third-party cybersecurity, including generative and agentic systems. The source establishes the consultation and its subjects. It does not establish that a final binding text had been issued on that date. ^[8]

The first operational principle is least privilege: an agent receives only the access required for the task; reading is separated from modification, testing from production, and recommendation from execution. Eloquence, and the label ‘autonomous agent,’ grant no access rights. Any transition to a high-impact action needs an explicit condition and documented human review or a technical control appropriate to the risk. ^[30]

The second principle is a record of the outcome. When reviewing a decision assisted by AI, an institution should be able to identify the tool version, relevant information sources, edits made, test applied, and person or function that authorized use. That does not mean preserving everything employees say. It means retaining the professionally necessary audit record under appropriate access and retention controls.

The third principle is cost per accepted outcome under an independent test. Cost includes task setup, tool operation, review, rework, and later defect handling. The outcome requires defined quality, not the mere completion of a file or closure of a ticket. This measure distinguishes more activity from more value.

11. Limitations and What This Paper Cannot Establish

The central case rests on Anthropic’s disclosure. The reviewed source set contains no independent examination of the account records or physical test. The cluster’s organizational identity cannot be verified; nor can the quality of the code or the fate of the

outputs after the ban. The counterfactual also remains unavailable: what the users would have completed, and how quickly, without Claude.

Regional sources are heterogeneous. UN reports, statements by governments and militaries, claims by armed groups, corporate reports, and journalism each have different access, interests, and standards of proof. This paper uses them to establish context, not to merge them into one homogeneous record. Conflicting accounts of some attacks and attributions remain unresolved.

This is an open-source narrative review, not a systematic review or meta-analysis. Its selection of examples may give more weight to cases that are well known or publishable. Provider reports are also shaped by detection capability and disclosure policy and offer no denominator from which prevalence can be estimated.

The paper does not provide a technical assessment of a weapons system or instructions for developing one. Details of design, targeting, interference, and system integration have been generalized because they are unnecessary to the argument and could make the analysis operational. The conclusions concern measurement of knowledge work, governance, and evidentiary limits rather than weapons engineering.

The open record may also contain publication bias. Companies disclose cases they detect and choose to publish; victims may not reveal every effect; successes and failures may not become visible at the same rate. The sources provide no common denominator for prevalence or cross-provider comparison. An increase in published reports alone therefore cannot support a quantitative trend claim.

12. A Civilian Research Agenda, Then the Remaining Question

The first research path concerns prior expertise. Groups with different levels of experience can be compared on safe civilian programming tasks, with random assignment of the tool, blind quality assessment, and measurement of total time. Does the tool most benefit people able to detect its errors, or does it genuinely narrow the expertise gap? A credible design must allow either result to appear.

The second path concerns burden shifted into verification. Record each stage separately: task preparation, draft production, review, testing, and rework. Then measure defects found before and after acceptance. The tool may turn out to accelerate both production and review. If so, the verification-tax hypothesis should change rather than being protected because it is elegant.

The third path concerns independence in automated review. Compare a reviewer that sees the writer's answer with one that begins from the specification and raw evidence, then measure defects found and defects both miss. Agent count should not be the only variable. Difference in evidence, authority to object, and a stopping criterion matter more.

A fourth path concerns the validity of results in Arabic and in local working environments. No assumption of weakness or superiority is needed in advance. Test the language, domain, and documents the institution actually uses. The purpose is not to manufacture a marketing ranking of models, but to learn where quality, cost, and omission risk change with language and context.

A fifth path can test governance itself. Compare teams using least privilege, a decision record, and independent testing with teams using the tool without those controls, on comparable safe civilian tasks. Measure late defects, recovery time, and audit cost, rather than satisfaction alone. Governance then becomes a measurable hypothesis instead of a handsome list of principles on the wall.

The Claude case leaves us with two things: a documented regional history of attacks and changing tools, and a new case whose evidence does not establish successful military effect. History makes the question urgent; it does not fill gaps in the source. Strategic analysis should identify what changed, what has not yet been shown to change, and what test could separate the two.

A risk need not be exaggerated to deserve attention, and confidence does not become protection because it is well phrased. Code, however convincing it looks on a screen, does not fly on its own. Hardware, data, testing, and responsibility stand between it and an outcome. That distance is where research and governance earn their value.

The Evidence Ladder: From Use to Effect

1	2	3	4
Observed use	Produced output	Independent verification	Accepted outcome

No step may be skipped: each transition needs a criterion and evidence independent of the party that produced the output.

References

1. Detecting and countering misuse of AI: September 2026
Anthropic
2. Users in Houthi-held Yemen tried to develop advanced weapons with AI, Anthropic says
Associated Press, September 11, 2026
3. GTIG AI Threat Tracker: From Prompting to Autonomy – The Evolution of Adversarial AI
Google Threat Intelligence Group, September 9, 2026
4. World Food Programme scales up response in Yemen as fighting displaces thousands
World Food Programme, September 17, 2026
5. World Food Programme ramps up response in Yemen as hunger fears escalate, official says
Reuters, September 21, 2026
7. 17 Iranians Charged With Conducting Massive Cyber Theft Campaign On Behalf Of The Islamic Revolutionary Guard Corps And Other Iranian Entities
U.S. Department of Justice, August 18, 2026
8. NCA Launches Public Consultation on AI Cybersecurity Guidelines
National Cybersecurity Authority, July 5, 2026
9. Final report of the Panel of Experts on Yemen
United Nations Security Council
10. United Nations Country Team in Yemen Annual Report 2021
United Nations in Yemen
11. Red Sea ballistic missile attacks trigger Asian interest in defences
Reuters, February 21, 2024
12. Final report of the Panel of Experts on Yemen
United Nations Security Council, October 11, 2024
16. Final report of the Panel of Experts on Yemen
United Nations Security Council, October 17, 2025

18. Shamoon 2: Return of the Disttrack Wiper
Unit 42, November 30, 2016
19. APT33: Insights into Iranian Cyber Espionage
Mandiant, September 20, 2017
20. Iranian Government-Sponsored Actors Conduct Cyber Operations Against Global Government and Commercial Networks
CISA, February 24, 2022
21. Incidents at Abqaiq and Khurais
Saudi Aramco, September 14, 2019
22. Security Council Press Statement on Terrorist Attacks in Abu Dhabi, United Arab Emirates
United Nations Security Council, January 21, 2022
23. Red Sea Attacks Disrupt Global Trade
International Monetary Fund, March 7, 2024
25. Why Do Multi-Agent LLM Systems Fail?
arXiv, March 17, 2025
26. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity
METR, July 10, 2025
27. We are Changing our Developer Productivity Experiment Design
METR, February 24, 2026
28. Task Substitution and Uplift
METR, May 8, 2026
29. Report of the Presidential Commission on the Space Shuttle Challenger Accident, Appendix F
NASA
30. The Protection of Information in Computer Systems
Proceedings of the IEEE