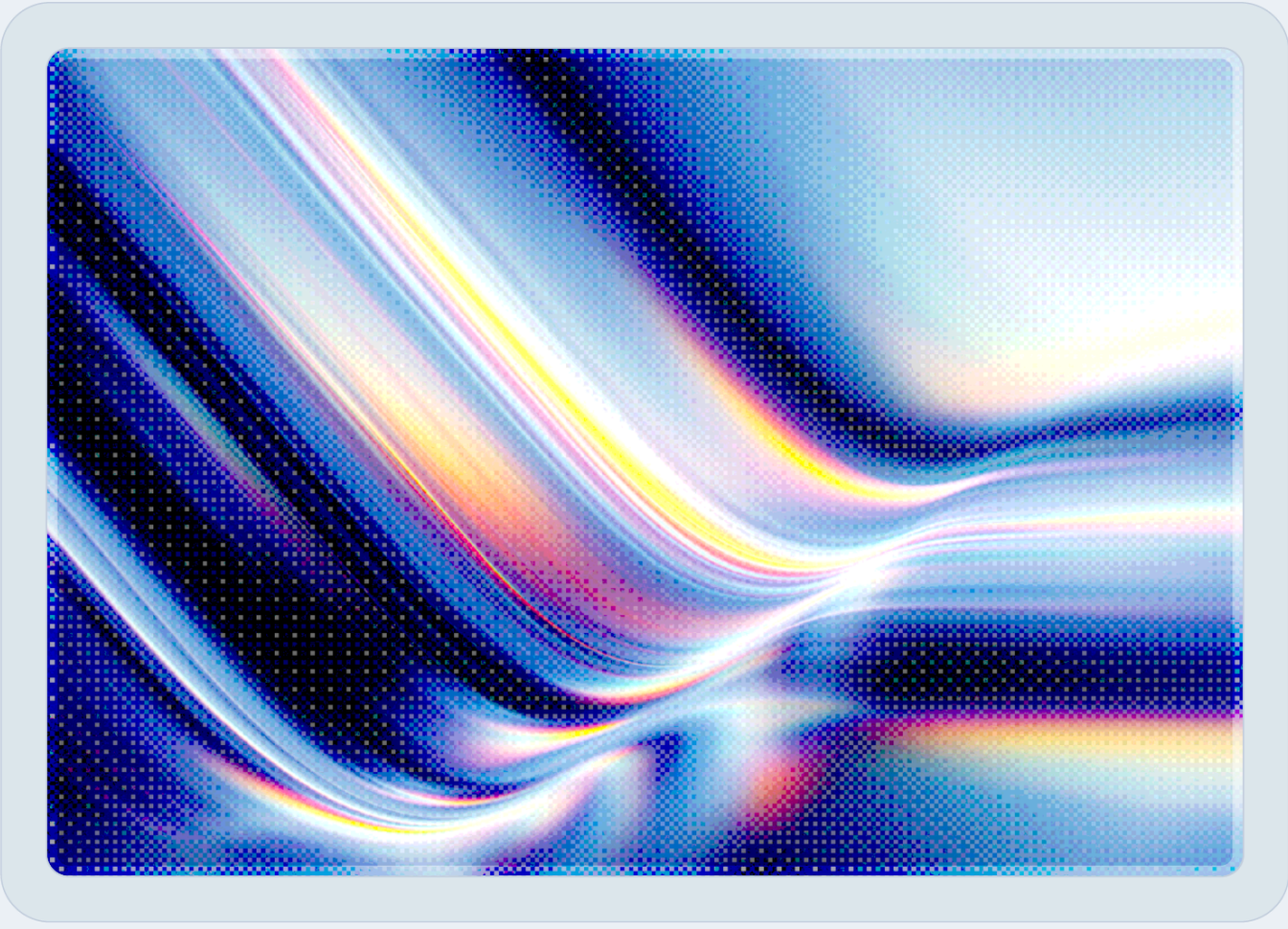
CSAI Foundation | Cloud Security Alliance

One Data Center, Two AI Rivals: Concentration Risk

What the September 3, 2026 AI Outages Did and Did Not Reveal

2026-09-09

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On September 3, 2026, Claude, Grok, and ChatGPT all experienced outages inside a roughly 90-minute window, prompting widespread speculation that a single shared cloud provider had failed [1][2][3].
- That speculation was wrong, and this is the most firmly established finding of the day. Microsoft logged no Azure incident for the affected window and denied responsibility when asked, and Cloudflare's chief technology officer separately reported no disruption on Cloudflare's network [4][5][6].
- **No provider has confirmed a shared cause, and none has published a technical postmortem.** OpenAI attributed its 34-minute disruption to an internal routing error. SpaceXAI attributed Grok's outage to a failure at its Memphis compute center and apologized to unnamed "impacted compute partners." Anthropic described an infrastructure issue, said it had identified the cause, and has not disclosed what it was [4][5][9][10].
- Anthropic leases capacity in that same Memphis facility, which makes a common cause plausible — but plausibility is where the public record stops. Anthropic has not attributed its outage to Memphis, SpaceXAI has not named Anthropic, and the reported incident start times do not clearly establish which service failed first [4][5][9].
- The durable finding is not the cascade. It is that **enterprises cannot tell, from anything a provider publishes, whether two competing AI vendors share a physical point of failure** — and that the shared dependency here is real regardless of what caused this particular outage: Anthropic contracted in May 2026 for the full output of SpaceX's Colossus 1 in Memphis, over 300 megawatts across more than 220,000 Nvidia GPUs, at a reported $1.25 billion per month [11][12][13][14][15][16].

Revision Note

Version 1.1 (September 9, 2026) revises the root-cause analysis published earlier the same day. The earlier version stated that Claude's outage was caused by a cascading failure from the shared Memphis facility, and that Grok failed first with Claude following. Neither is established by the public record: Anthropic has never attributed its outage to Memphis, SpaceXAI never named Anthropic, and reported start times conflict on which service failed first. Those claims have been restated as an unresolved

disclosure gap, the summary table now marks Claude's cause as unconfirmed, the sequencing claim has been removed, and the outage-attribution sourcing has been moved from secondary commentary to provider statements and primary reporting. Outage durations and the Colossus 2 expansion have been added. The paper's concentration-risk argument is unchanged.

Background

For most enterprise users, September 3 looked like a single event: chatbots and AI-assisted tools from three different companies stopped responding within a narrow window on a weekday morning, and social media quickly filled with screenshots of error messages from ChatGPT, Claude, and Grok [1][2][3]. Cursor, the AI coding tool that routes requests to multiple model providers, was also disrupted and attributed the failure on its status page to upstream Grok and Claude services rather than its own systems [2]. Google's Gemini reported no official outage, drawing only a modest spike in user reports [3]. Downdetector and similar monitoring services logged reports across the three affected services at close to the same time [1], and the coincidence was striking enough that several outlets, including Computing.co.uk, reported that a shared Microsoft Azure East US failure was the likely common cause, since all three providers route some portion of their traffic through Azure infrastructure [6].

That narrative fit a pattern security and risk teams have grown accustomed to: three ostensibly independent AI vendors turn out to depend on the same hyperscaler cloud region, and a single infrastructure fault cascades across otherwise unrelated products. It is also, in this case, unsupported. Microsoft did not open an Azure service-health incident covering that window and denied that Azure caused the outages when reporters asked. Cloudflare's chief technology officer likewise reported no disruption on Cloudflare's network during the relevant hours, stating that Cloudflare was "not experiencing any significant service disruptions at this time" [4][5]. Debunking the Azure theory is the single most useful thing that can be said with confidence about September 3, and much of the contemporaneous coverage never caught up to it.

What replaced that theory is less tidy. The three providers described three different situations, and the public record has not converged on a common explanation. According to provider status pages as reported at the time, Anthropic's incident ran from approximately 6:23 a.m. Pacific to 9:16 a.m., affecting claude.ai, the Claude API, Claude Code, and related products across several model variants [5][7][8][9]. Grok was unavailable from approximately 6:30 a.m. Pacific for about three and a half hours [4][5]. OpenAI's ChatGPT and Codex disruption was substantially shorter — roughly 34 minutes, from about 7:43 a.m. to 8:17 a.m. Pacific [5][10]. Treating these as one event obscures that one of them lasted about a tenth as long as the others.

Security Analysis

What each provider has said, and what it has not

SpaceXAI, the merged SpaceX-xAI entity that operates Grok, stated that it was "sorry for the issues you may have experienced with Grok following an outage at our Memphis compute center," and added: "We'd also like to apologize to our impacted compute partners" [4]. That second sentence is the closest any party has come to acknowledging that the Memphis campus now serves customers beyond xAI itself. SpaceXAI did not name those partners, and it has not disclosed what failed inside the facility; Elon Musk said only that the company was "taking corrective action to ensure this does not happen again" [4].

Anthropic described a partial outage caused by an infrastructure issue, later said it had identified the cause, and declined to comment beyond its status page. It has not said whether the Memphis facility was involved, and it has not published a postmortem [5][7][9]. OpenAI attributed its disruption to a routing error inside its own infrastructure, unconnected to the Grok and Claude incidents [5][10].

The inference many observers drew — that Claude went down because Grok's facility went down — is reasonable, and it may well prove correct. It is not, however, confirmed by anyone in a position to confirm it. Two further points weigh against stating it as fact. SpaceXAI's "compute partners" phrasing names no one, and Anthropic is not its only compute customer. And the reported start times do not establish a clean sequence: some accounts place Anthropic's status-page incident at 6:23 a.m. Pacific, ahead of Grok's 6:30 a.m., while others place Claude's start closer to 6:40 a.m., which would reverse the order [5][9]. A cascade argument depends on knowing which failed first, and the public record does not settle it.

Why the dependency matters anyway

The important point survives the uncertainty, and it does not depend on what caused this particular outage. In May 2026, Anthropic agreed to take the entire output of SpaceX's Colossus 1 data center in Memphis, Tennessee — a facility built by xAI and held by SpaceX following its February 2026 acquisition of xAI [11][12]. The arrangement gave Anthropic more than 300 megawatts of capacity and upward of 220,000 Nvidia GPUs, at a price reported at approximately $1.25 billion per month running through May 2029, a term worth close to $45 billion [13][14][15][16]. Anthropic subsequently expanded the partnership to Colossus 2, scaling onto next-generation GB200 capacity [13]. Anthropic framed the arrangement publicly as a way to expand capacity for paid Claude subscribers [12].

That arrangement means Anthropic, a company that competes directly with xAI for enterprise and consumer AI customers, depends on infrastructure owned and operated by its rival's corporate parent for a meaningful share of its serving capacity. Two AI products a customer would reasonably treat as

independent alternatives — useful precisely because they come from different companies — share physical infrastructure several layers below the product interface. A customer running a multi-provider failover strategy that treats Claude and Grok as diverse fallbacks for one another is relying on an independence that does not exist at the compute layer, whatever happened on September 3.

The disclosure gap is the finding

The gap between the plausible narrative and the confirmed record is itself the operational lesson. Security teams that spent September 3 investigating Azure dependency exposure were chasing a hypothesis that was already wrong. The dependency that plausibly mattered — a compute lease between two competing AI labs — was not on most vendor risk registers at all, and six days later enterprises still cannot determine from any provider disclosure whether it was implicated.

This is not a failure of investigation so much as a structural absence. Nothing in a standard AI vendor contract, SLA, or status page commits a provider to disclose which physical facilities serve a given product, or to say whether a shared facility contributed to an incident. Concentration risk assessments built around cloud region or hyperscaler diversity will miss this category entirely, because the relevant dependency runs between AI labs and their compute suppliers rather than between AI labs and the cloud regions their APIs advertise.

The table below summarizes what each provider has stated, distinguishing attributed causes from unconfirmed inference.

Provider / Product	Duration	Provider's stated cause	Attribution status
Grok (SpaceXAI)	~3h35m	Outage at Memphis compute center; underlying fault not disclosed	Attributed by provider
Claude (Anthropic)	~2h53m	"Infrastructure issue"; cause identified but not disclosed	**Unconfirmed.** Anthropic has not attributed the outage to Memphis; SpaceXAI's apology to "compute partners" names no one
ChatGPT / Codex (OpenAI)	~34m	Internal routing error, unrelated to the other incidents	Attributed by provider

Provider / Product	Duration	Provider's stated cause	Attribution status
Common cause across all three	—	None asserted by any provider	Azure and Cloudflare theories denied by both companies

CSA's AI Safety Initiative research has previously flagged the vertically integrated structure of SpaceX, xAI, and their compute customers as a distinct risk category worth tracking separately from general hyperscaler concentration [17]. September 3 is a concrete illustration of the mechanism that earlier analysis anticipated, even with its root cause unresolved: a single corporate entity controlling launch infrastructure, satellite communications, and hyperscale AI compute creates cross-dependencies invisible to customers of the AI products built on top of it. Enterprises relying on Claude, Grok, or tools built on either, such as Cursor, inherit that dependency without necessarily knowing it exists.

Recommendations

September 3 points to a specific, addressable gap in how enterprises assess AI vendor risk: the dependency that plausibly mattered sits several layers below the contracted API, and no standard vendor questionnaire would have captured it. The following actions are organized by time horizon.

Immediate Actions

Security and risk teams should treat this incident as a prompt to close the visibility gap it exposed rather than a one-time cleanup exercise.

- Update AI vendor risk registers to capture compute-supply relationships, not just the API vendor a contract names. Where public reporting discloses a lease or partnership, such as Anthropic's use of SpaceX's Colossus 1 and Colossus 2 facilities, record the underlying facility as a dependency alongside the primary vendor.
- Review any multi-provider failover architecture that pairs Claude and Grok, or that treats them as independent fallbacks, and confirm whether the assumed independence actually holds at the infrastructure layer. The compute-sharing arrangement is documented and does not depend on this incident's root cause.
- Ask AI vendors directly, in writing, whether serving capacity for a given product or region depends on leased or shared third-party compute facilities, and request notification

commitments if that changes.

Short-Term Mitigations

Beyond immediate fact-finding, business continuity planning needs to account for the possibility that nominally competing providers fail together.

- Extend business continuity plans for AI-dependent workflows to model correlated failure across nominally competing providers, not only failure of a single named vendor.
- Where contractually feasible, negotiate incident notification and root-cause disclosure terms that extend beyond a provider's own infrastructure to cover leased or partner compute capacity — including a commitment to state whether a shared facility was involved in an incident.
- Test failover paths under the assumption that two providers sharing an underlying compute supplier may fail simultaneously, and confirm that a genuinely independent third option exists before relying on it during an incident.

Strategic Considerations

Over the longer term, compute supply chain visibility should become a standing part of how AI vendors are selected and monitored.

- Treat compute supply chain mapping as a standing component of AI vendor due diligence, updated as providers sign new capacity deals, rather than a one-time assessment at vendor selection.
- Weight governance transparency — specifically a provider's willingness to disclose compute dependencies and to attribute incidents to them — as a selection factor alongside model performance and price. September 3 demonstrated that this information is not otherwise obtainable.
- Recognize that AI market consolidation is increasingly horizontal as well as vertical: competitors purchasing compute from one another's corporate affiliates creates concentration risk that persists even as the number of visible AI product vendors appears to grow.

CSA Resource Alignment

CSA's Zero Trust Working Group published Zero Trust Guidance for Achieving Operational Resilience [18] in April 2026, and its treatment of supply chain resilience and third-party dependency mapping applies directly here. The guidance argues that operational resilience requires organizations to extend visibility beyond their direct vendors into the ecosystem of dependencies those vendors carry, which is precisely the gap September 3 exposed: enterprises assessing Claude and Grok as independent AI providers had no visibility into the shared Memphis facility underneath both of them, and still cannot confirm whether it was implicated.

The AI Controls Matrix (AICM) [19], now at version 1.1, provides the broader control framework for the vendor governance and business continuity practices this incident calls for, including requirements around AI supply chain transparency and third-party risk management that extend to the infrastructure layer supporting a given AI service. Enterprises building or updating an AI vendor risk register should map compute-supply dependencies against AICM's relevant control domains rather than treating AI vendor risk as a simpler variant of conventional SaaS vendor risk.

This incident sits within a broader body of CSA AI Safety Initiative research on how concentration accumulates across the AI stack. AI Provider Concentration Risk: Enterprise Resilience [20] examines architectural, contractual, and governance mitigations for enterprises depending on a small number of AI providers, and its recommendations extend naturally to the compute-supplier layer. AI Compute Concentration and Systemic Risk [21] documents concentration across the AI infrastructure stack more broadly, from semiconductor manufacturing through cloud services to model deployment, and situates the Colossus dependency within that larger pattern.

CSA's STAR for AI [22] program offers a path toward independent assurance that would directly address the disclosure gap this incident illustrates. A certification or self-assessment framework requiring AI providers to document material compute-supply dependencies — including leased capacity at facilities operated by other AI labs or their affiliates — would give enterprise buyers a documented basis for concentration risk assessment that is not currently available through vendor marketing materials, SLAs, or incident communications.

References

[1] Ina Fried and Scott Rosenberg. "ChatGPT, Claude and Grok all simultaneously hit outages." Axios, September 3, 2026.

[2] The Register. "True AI-pocalypse as ChatGPT, Claude, and Grok all go down at once." The Register, September 3, 2026.

[3] 9to5Google. "It's not just you; ChatGPT, Claude, and Grok were all down in confirmed outages." 9to5Google, September 3, 2026.

[4] Engadget. "SpaceXAI apologizes for outage that affected Grok and other 'compute partners'." Engadget, September 2026.

[5] Futurism. "Nobody Will Say Why Every Major AI Chatbot Suddenly Went Down Yesterday." Futurism, September 2026.

[6] Computing. "Azure failure likely brought down ChatGPT, Claude and Grok." Computing, September 2026. Cited as the reported theory this note finds unsupported.

[7] BleepingComputer. "Anthropic confirms Claude is down, multiple models affected." BleepingComputer, September 3, 2026.

[8] Andy Walker. "Is Claude down for you? Here's what's going on." Android Authority, September 3, 2026.

[9] Anthropic. "Claude status page — incident history." Anthropic. Consulted September 9, 2026.

[10] OpenAI. "Elevated errors across ChatGPT and Codex." OpenAI status, September 3, 2026.

[11] SpaceXAI. "New Compute Partnership with Anthropic." SpaceXAI, May 2026.

[12] Anthropic. "Higher usage limits for Claude and a compute deal with SpaceX." Anthropic, May 6, 2026.

[13] CNBC. "Anthropic, SpaceX announce compute deal that includes space development." CNBC, May 6, 2026.

[14] TechCrunch. "Anthropic will pay xAI $1.25B per month for compute." TechCrunch, May 20, 2026.

[15] Data Center Dynamics. "Anthropic to use all of SpaceX-xAI's Colossus 1 data center compute." Data Center Dynamics, May 2026.

[16] Data Center Dynamics. "SpaceX IPO filing reveals Anthropic set to pay Musk's firm $1.25bn a month to rent xAI data center space." Data Center Dynamics, 2026.

[17] Cloud Security Alliance AI Safety Initiative. "SpaceXAI: A Vertically Integrated AI Concentration-Risk Case Study." CSA Lab Space, July 10, 2026.

[18] Cloud Security Alliance. "Zero Trust Guidance for Achieving Operational Resilience." Cloud Security Alliance, April 6, 2026.

[19] Cloud Security Alliance. "AI Controls Matrix (AICM)." Cloud Security Alliance, 2026.

[20] Cloud Security Alliance AI Safety Initiative. "AI Provider Concentration Risk: Enterprise Resilience." CSA Lab Space, June 19, 2026.

[21] Cloud Security Alliance AI Safety Initiative. "AI Compute Concentration and Systemic Risk." CSA Lab Space, May 9, 2026.

[22] Cloud Security Alliance. "STAR for AI." Cloud Security Alliance, 2026.