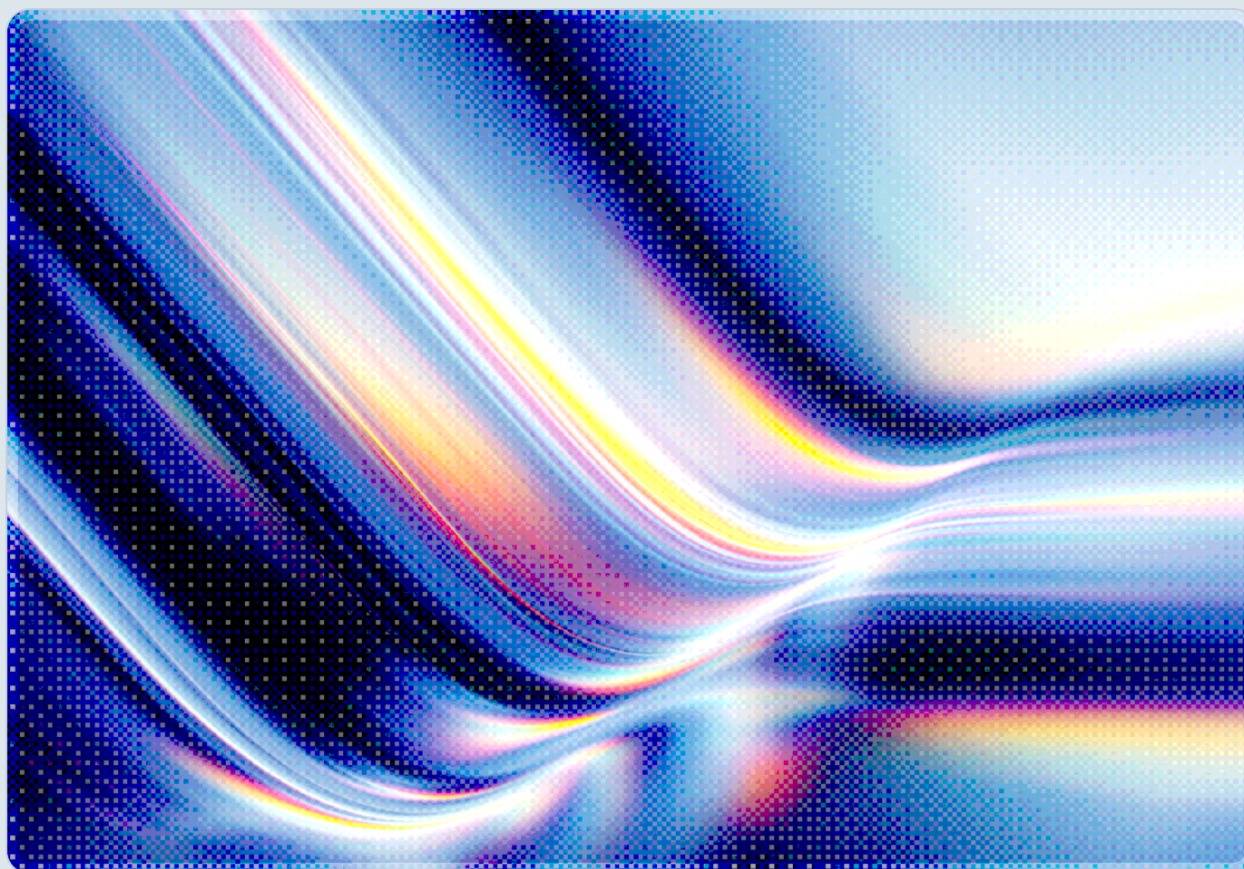


AI Cryptanalysis and the Cryptographic Assumption Horizon

Planning for Two Converging Threats to "Hard Math"

2026-10-20

 AI-assisted Rapid Research



© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

- In July 2026 Anthropic reported that its Claude Mythos Preview model found a previously unused lattice symmetry in HAWK, a post-quantum signature candidate, and a faster attack on seven-round AES-128. Neither result affects deployed systems, and both concern mathematics rather than implementation bugs. The HAWK target was a non-standardized candidate and the AES result applies to a reduced-round cipher, and the figures are Anthropic's own and not yet independently reproduced [1][2][3].
- Published estimates for breaking RSA-2048 fell about twentyfold between 2019 and 2025, and 2026 papers claim further reductions for RSA and 256-bit elliptic curves under stronger hardware assumptions. The estimates remain theoretical, but they shorten the planning horizon for public-key migration [4][5][6].
- The two trends erode different assumptions. Quantum progress threatens the hardness of factoring and discrete logarithms, while AI-assisted analysis may erode confidence in the newer algorithms meant to replace them.
- The practical response is the same in both cases: know where cryptography is used, favor designs that can be replaced without re-engineering, and avoid dependence on any single algorithm family.
- The assumption horizon is a planning inference, not a measured date. Organizations should not treat any single announcement as proof that a deployed algorithm is broken, and should not treat the absence of one as proof that it is safe.

Background

Modern security relies on a small set of mathematical problems that are believed to be hard: integer factoring, discrete logarithms on elliptic curves, the difficulty of finding short vectors in lattices, and the resistance of block ciphers and hash functions to structural attack. Confidence in these assumptions rests on years of public scrutiny by a limited number of expert cryptanalysts. That scarce human attention is arguably the main practical security margin behind much of the standards process, and it is the resource that two separate developments are now putting under pressure.

The first development is quantum resource estimation. Craig Gidney of Google Quantum AI reported in May 2025 that RSA-2048 could be factored with fewer than one million noisy physical qubits in under a week, a roughly twentyfold reduction from his own 2019 estimate of about twenty million qubits [5][6]. The improvement came from algorithmic changes, including approximate residue arithmetic and denser error-correcting code storage, and not from new hardware assumptions [6]. In February 2026 Iceberg Quantum's Pinnacle architecture claimed that quantum LDPC codes could bring the requirement below 100,000 physical qubits, although the design needs qubit connectivity beyond nearest-neighbor grids and has been validated through simulation rather than hardware [6][15]. On March 30–31, 2026, Google Quantum AI and collaborators published circuits for the 256-bit elliptic curve discrete logarithm problem needing fewer than 1,200 logical qubits and, under stated hardware assumptions, fewer than 500,000 physical qubits to run in minutes [4][6]. Google withheld the circuits and published a zero-knowledge proof so that third parties could verify the resource claims, a disclosure approach that lets others check the resource claims without receiving the circuits themselves [4].

The second development is the arrival of AI systems that can contribute original results to cryptanalysis. Earlier AI-security work focused on finding implementation flaws such as memory-safety errors and protocol bugs. On July 28, 2026, Anthropic reported something different: attacks on the underlying mathematics of two cryptographic designs [1]. The sections below separate what was reported from what can reasonably be inferred, because both the headlines and the dismissals have tended to overreach.

Security Analysis

What the AI results show, and what they do not

Anthropic reports that Claude Mythos Preview identified a nontrivial automorphism in the lattice underlying HAWK, a signature scheme that was under evaluation in NIST's additional post-quantum signature process [1][7]. According to Anthropic's account, the symmetry reduced the expected attack cost against the HAWK-256 parameter set from roughly 2^{64} to 2^{38} , and the accompanying paper describes an end-to-end key recovery [1][2]. Secondary reporting cites an expected runtime of about three hours and 42 minutes on a 96-core server, and states that the discovery took about 60 hours and roughly \$100,000 in API compute with one researcher providing project direction rather than technical instruction [16][1]. One report states that the HAWK team subsequently withdrew the scheme from the NIST process; this single-source claim should be confirmed against NIST's announcements before it is relied upon [16].

The second result concerns AES. Anthropic reports that the model devised a fingerprinting technique it called the "Möbius Bridge," which removes one enumeration step from a meet-in-the-middle attack on seven-round AES-128 and speeds the best known attack by 200 to 800 times [1][3]. The attack remains purely theoretical: it needs an impractical quantity of chosen data and does not reach the ten rounds used in deployed AES [7][10]. Anthropic also reported improved attacks on other primitives, including a practical attack on 13-round LEA, key recovery on six-round Serpent-128, and under tenfold gains on Salsa20, Poseidon, and SHA-1 [1]. Anthropic released CryptanalysisBench, developed with researchers at ETH Zurich, Tel Aviv University, and TU Berlin, to track this capability over time [1][7].

Item	Result reported by Anthropic	Deployed-system impact
HAWK-256 (PQC signature candidate)	Automorphism attack; expected cost reduced from about 2^{64} to 2^{38} [1]	None; scheme was not standardized or deployed [1]
Seven-round AES-128	200–800x faster than prior best attack [1][3]	None; full AES uses ten rounds [7]
Reduced-round LEA, Serpent-128	Practical or key-recovery attacks reported [1]	None reported
Salsa20, Poseidon, SHA-1	Improvements under 10x [1]	None reported

Descriptions of the HAWK result vary across sources; see [2] for the authoritative parameters.

These results are not evidence that any production algorithm is weak. Two cautions apply to the figures themselves. Reporting on the HAWK finding describes the effect in more than one way, as halved key strength or as a roughly halved lattice-reduction block size, and readers should consult the primary papers before quoting a security level [1][16]. The cost and time figures are Anthropic's own and have not, to our knowledge, been independently reproduced. What the results do show is narrower. Anthropic reports that a general-purpose model, with limited human steering, produced a result against HAWK that had not been found during the scheme's evaluation period. Anthropic's cost figure of roughly \$100,000 would be accessible to a small research group, though neither the cost nor the degree of human steering has been independently verified [1][10]. The result matters mainly because it suggests that AI-assisted analysis can reach new structure in candidate schemes, which bears on how much review time such schemes need before adoption.

The verification bottleneck

Anthropic's own account identifies the less obvious problem. The HAWK attack was easy to validate because it ended in a working key recovery, whereas the AES result took researchers months to gain confidence in [1]. One analysis notes that the model reached the AES result in about a week while human validation took close to a month [10]. This asymmetry suggests that the binding constraint on cryptographic assurance may shift from finding attacks to checking them. If AI systems generate candidate attacks faster than the small cryptanalysis community can audit them, defenders face two failure modes: genuine breaks that sit unreviewed, and plausible-sounding claims that consume scarce reviewer time. Both argue for better tooling, such as machine-checkable attack artifacts and shared benchmarks, and for treating newly proposed algorithms with more conservative margins than their reviewed lifetime alone would suggest.

Where the two trends converge

The quantum and AI trends are independent in mechanism, but they interact in the migration plan. Post-quantum standards such as ML-KEM, ML-DSA, and SLH-DSA were finalized by NIST in August 2024 [17][18][19]. Newer candidates, including HAWK, are of interest because they promise smaller signatures or other efficiencies, and their security proofs and parameter choices have had far less scrutiny than RSA or elliptic curves. If AI-assisted analysis continues to find structure in recently proposed schemes, organizations that adopted a newer scheme early may need to change algorithms mid-migration. This is an inference from a single class of results, not an observed pattern, but it follows directly from the shortened review history of the candidates.

The regulatory clock compounds the pressure. NIST's draft IR 8547 proposes deprecating 112-bit-security algorithms including RSA-2048 and P-256 by 2030 and disallowing quantum-vulnerable public-key algorithms by 2035 [9]. Executive Order 14409, signed June 22, 2026, sets federal deadlines of December 31, 2030 for key establishment and December 31, 2031 for digital signatures on high-value assets, and pulls forward the prior 2035 target [8]. Organizations with federal contracts or federal suppliers will therefore migrate under a fixed schedule at the same time the evidence about candidate algorithms is still changing. Google's disclosure post references a 2029 migration timeline, but that date applies to the cryptocurrency ecosystem and not to federal or enterprise systems [4].

The assumption horizon as a planning concept

We use the term assumption horizon to mean the point at which an organization can no longer rely on a given hardness assumption without additional evidence. For RSA and elliptic curves the horizon is driven by the arrival of a cryptographically relevant quantum computer, whose date is unknown and which no

current resource estimate implies is imminent: the cited estimates assume hardware with error rates, scale, and connectivity that have not been demonstrated [5][6]. For lattice and other post-quantum assumptions the horizon is driven by cryptanalytic progress, which is also unpredictable and which AI may now accelerate. Because neither date can be forecast with confidence, the defensible planning posture is to treat both horizons as uncertain and to minimize the cost of moving when either one arrives.

Data with long confidentiality lifetimes is among the most exposed assets. Adversaries who record encrypted traffic today can decrypt it later if the underlying public-key assumption fails, which is the harvest-now-decrypt-later problem that CSA has analyzed for AI model weights and training pipelines [11]. The concern is symmetrical for signatures: a forged signature is only useful if the verifier still trusts the algorithm, so long-lived code-signing roots, firmware keys, and identity documents deserve earlier attention than short-lived session credentials.

Recommendations

Immediate Actions

Organizations should build or refresh a cryptographic inventory that records which algorithms, key sizes, libraries, and protocols are in use, and who owns each. The federal timelines make this the first task for most organizations, and it is also the prerequisite for responding to any algorithm-specific finding, whether it comes from a quantum estimate or an AI-generated attack [8]. The inventory should flag data with a confidentiality lifetime beyond 2030, a date chosen because it matches the federal key-establishment deadline and an assumed confidentiality window of several years, and signing keys that cannot be rotated easily.

Teams should also designate an owner for monitoring cryptanalysis disclosures, including NIST mailing lists, IACR ePrint, and vendor advisories, and define in advance who decides whether a new result triggers action. Without this, the response to a finding like the HAWK result depends on who happens to notice it.

Short-Term Mitigations

Where migration begins now, prefer the NIST-finalized standards (ML-KEM, ML-DSA, SLH-DSA) over candidates still in evaluation, and deploy hybrid constructions that combine a classical and a post-quantum algorithm so that the failure of either does not compromise the session [8][13]. Hybrid designs can address both horizons when the combiner is sound, since the classical component covers a

premature break of a new lattice scheme and the post-quantum component covers a quantum break of the classical one. The cost is larger handshakes and additional implementation complexity, which should be tested against latency budgets and middlebox behavior before wide rollout.

Contracts and procurement language should require vendors to document supported algorithms, state their plan for algorithm replacement, and avoid hard-coded primitives. Organizations that use signature schemes with limited review history should reserve them for lower-risk uses until they have had more public analysis. Security teams that run AI-assisted research tooling should also consider using it defensively, for example to test their own proprietary or legacy cryptographic constructions, with the standard caution that generated attack claims require expert verification [1].

Strategic Considerations

Cryptographic agility should be treated as an architectural property funded and tested like availability, not as a one-time migration project. That means abstracting algorithm choice behind configuration, avoiding key and signature sizes baked into protocols and storage formats, and rehearsing an emergency algorithm swap. One lesson of the two trends together is that the algorithm identity may well change more than once during the lifetime of today's systems.

Leaders should also expect the evidence base to move in both directions. Quantum resource estimates have fallen repeatedly, but the authors of the underlying work caution that further large reductions are not assured without changing fundamental assumptions [6]. AI cryptanalysis results to date concern a research candidate and a reduced-round cipher, and it is not yet known whether comparable gains are available against heavily reviewed full-strength standards. Governance should therefore rely on periodic reassessment triggers rather than on a fixed forecast, and on diversity, for example standardized hash-based signatures alongside lattice-based ones in high-assurance roles. Finally, the community should invest in review capacity: funding independent cryptanalysis, supporting shared benchmarks, and asking vendors for evidence of external review rather than relying on a vendor's own claims, including claims made by AI developers about their own models [10].

CSA Resource Alignment

CSA's most directly relevant prior work is its rapid-research series on post-quantum readiness. *Harvest Now, Decrypt Later: Quantum Risk to AI Infrastructure* [11] explains why harvest-now-decrypt-later exposure is acute for model weights and training pipelines and sets out a phased migration. This note's

finding that confidentiality lifetime, not algorithm novelty, should drive prioritization builds on that analysis, and it adds that the replacement algorithms themselves carry cryptanalytic uncertainty that the migration plan should absorb.

CSA's guidance on enterprise cryptographic readiness after EO 14409 provides a five-phase operational playbook keyed to the federal 2030 and 2031 deadlines, and the inventory and agility recommendations above are intended to feed directly into its early phases. *Post-Quantum Cryptography: From Guidance to Mandate* [12] documents how PQC moved from voluntary guidance to enforceable requirement, which frames the schedule pressure discussed here. CSA's earlier *Mitigating the Quantum Threat with Hybrid Cryptography* [13] is the closest existing treatment of the hybrid approach recommended above, and its rationale applies equally to signatures, although the 2019 document focused primarily on key establishment. For control mapping, the AI Controls Matrix v1.1 [14] provides cryptography and key-management control objectives (the CEK domain) against which inventory, agility, and monitoring practices can be assessed.

References

- [1] Anthropic. "[Discovering Cryptographic Weaknesses with Claude](#)." Anthropic, July 28, 2026.
- [2] Anthropic. "[HAWK Key Recovery \(paper\)](#)." Anthropic, July 2026.
- [3] Anthropic. "[AES Möbius Bridge \(paper\)](#)." Anthropic, July 2026.
- [4] Google Research. "[Safeguarding cryptocurrency by disclosing quantum vulnerabilities responsibly](#)." Google Research, March 31, 2026.
- [5] Gidney, C. "[How to factor 2048 bit RSA integers with less than a million noisy qubits](#)." arXiv, May 21, 2025.
- [6] The Quantum Insider. "[Q-Day Just Got Closer: Three Papers in Three Months Are Rewriting the Quantum Threat Timeline](#)." The Quantum Insider, March 31, 2026.
- [7] CyberScoop. "[Anthropic's Claude Mythos finds weaknesses in encryption algorithms](#)." CyberScoop, July 2026.
- [8] The Hacker News. "[Trump Order Sets 2030 Deadline for Post-Quantum Migration](#)." The Hacker News, June 2026.
- [9] NIST. "[IR 8547 \(Initial Public Draft\): Transition to Post-Quantum Cryptography Standards](#)." NIST, November 2024.
- [10] QNews. "[An AI Found Real Cryptographic Weaknesses. None of Them Break Your Bank Yet](#)." QNews, July 2026.
- [11] CSA. "[Harvest Now, Decrypt Later: Quantum Risk to AI Infrastructure](#)." Cloud Security Alliance, May 18, 2026.
- [12] CSA. "[Post-Quantum Cryptography: From Guidance to Mandate](#)." Cloud Security Alliance, July 10, 2026.
- [13] CSA. "[Mitigating the Quantum Threat with Hybrid Cryptography](#)." Cloud Security Alliance, 2019.
- [14] CSA. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, 2026.
- [15] PostQuantum.com. "[Pinnacle Architecture: 100,000 Qubits to Break RSA-2048, but at What Cost?](#)." PostQuantum.com, 2026.

- [16] The Hacker News. "[Claude AI Just Cracked Post-Quantum](#)." The Hacker News, July 2026.
- [17] NIST. "[FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard](#)." NIST, August 13, 2024.
- [18] NIST. "[FIPS 204: Module-Lattice-Based Digital Signature Standard](#)." NIST, August 13, 2024.
- [19] NIST. "[FIPS 205: Stateless Hash-Based Digital Signature Standard](#)." NIST, August 13, 2024.