● Ainary

AR-036 Update to AR-001 Confidence: 71%

State of AI Agent Trust
Q1 2026 Update

Two weeks after AR-001 shipped, the trust race got harder to win. 341 malicious skills on ClawHub. New benchmarks for auditing agent reasoning. The creator of the most popular personal AI agent just joined OpenAI. Here's what it means for the thesis.

February 17, 2026

v3.0 — Supplement to AR-001 v2.3

Florian Ziesche · Ainary Ventures

CONTENTS

1. What Changed Since AR-001 UPDATE

AR-001 (*State of AI Agent Trust 2026*, v2.3) shipped on February 15, 2026. It introduced the Trust Race Model — the thesis that agent capability grows exponentially while governance updates linearly, creating a structurally widening "ungoverned capability zone." We scored overall confidence at 73%.

Forty-eight hours later, four developments landed that directly affect the report's core claims:

341

malicious skills found on ClawHub (now 824+)

Koi Security / ClawHavoc [S1]

+5%

accuracy from AgentAuditor over majority vote

arXiv:2602.09341 [S2]

81%

of orgs plan complex multi-step agent workflows in 2026

Anthropic State of AI Agents [S5]

- **ClawHavoc supply chain attack:** Koi Security audited all 2,857 skills on ClawHub and found 341 were malicious — 335 from a single campaign distributing the Atomic Stealer (AMOS) infostealer via fake "prerequisites." The number has since grown to 824+.[S1] E
- **AgentAuditor:** New framework for auditing multi-agent reasoning trees yields up to 5% absolute accuracy improvement over majority vote and 3% over LLM-as-Judge. Introduces Anti-Consensus Preference Optimization (ACPO).[S2] E
- **AIRS-Bench:** First standardized benchmark for AI research science agents — 20 tasks from state-of-the-art ML papers. Even when agents surpass human benchmarks, they don't reach theoretical ceilings.[S3] E
- **Peter Steinberger joins OpenAI:** OpenClaw's creator hired to "drive the next generation of personal agents" (Sam Altman). OpenClaw moves to a community-run foundation backed by OpenAI.[S4] E

- **Anthropic's enterprise data:** 81% of organizations plan to expand into more complex agent use cases in 2026, with 57% already deploying multi-step agent workflows.[S5] E

> NET EFFECT ON AR-001 THESIS
>
> The Trust Race Model is **reinforced, not invalidated**. ClawHavoc demonstrates that the ungoverned capability zone isn't just theoretical — it's being actively exploited. New benchmarks show measurement is improving, which is the prerequisite for closing the gap. But deployment pressure (81% expanding) is accelerating faster than governance. Overall confidence adjusts down slightly from 73% to 71%. Rationale in Section 7.

2. Trust Race Update: The Gap Is Widening

72%

(Confidence: Medium-High)

In AR-001, we argued that deployment pressure pushes unreliable agents into production faster than governance can catch up. Anthropic's own enterprise data now quantifies the pressure: 81% of organizations plan to expand into more complex agent use cases in 2026.

In our original report (AR-001, §4), we documented the 51-percentage-point chasm between deployment (57%) and trust (6%). The new Anthropic data adds a velocity dimension: it's not just that enterprises deploy without trust — they're **actively planning to accelerate.**[S5] E

Consider the numbers together. 57% of organizations already deploy multi-step agent workflows[S5]. 81% plan to expand into more complex use cases[S5]. But still only 6% fully trust agents for core processes (AR-001, §4, citing HBR[AR-001, ref 1]). Only 12% have governance controls in place[AR-001, ref 1]. The deployment-trust gap isn't closing — it's widening under acceleration. I

Exhibit 1: Trust Race — New Data Points (Feb 2026)

METRIC	AR-001 (FEB 15)	THIS UPDATE (FEB 17)	DIRECTION
Deployment pressure	86% plan to increase investment (HBR)	81% plan complex multi-step workflows (Anthropic)	Reinforced ↑
Trust level	6% full trust (HBR)	No new data	Unchanged
Active exploitation of trust gap	Theoretical (§8–9 cascade)	341–824 malicious skills in the wild	Now empirical ↑
Measurement capability	TRiSM proposed metrics, unvalidated	AgentAuditor: +5% over majority vote; AIRS-Bench: 20-task suite	Improving ↑
Ecosystem governance	ISO 42001, NIST AI RMF, EU AI Act	ClawHub: no vetting process for skills	Worse at edges ↓

Sources: Anthropic [S5], Koi Security [S1], arXiv:2602.09341 [S2], arXiv:2602.06855 [S3]. AR-001 references per original numbering.

The Trust Race Model's four components (Capability Velocity, Reliability Floor, Governance Tempo, Deployment Pressure — AR-001, §8) all moved in the direction the model predicted: capability improving via new benchmarks, governance still lagging, deployment pressure increasing, and the ungoverned zone now actively exploited. This is not confirmation bias — it's two weeks of data. But it's consistent. I

SO WHAT?

The 81% expansion figure changes the urgency calculus from AR-001's recommendations. In §13 of the original, we gave a 90-day plan. The ClawHub crisis suggests Month 1 ("Measure the Gap") should now include a **supply chain audit** of all installed agent skills and extensions — not just a deployment inventory.

3. Agent Security: The ClawHub Crisis 85%

(Confidence: High — Multiple independent sources)

In AR-001 (§9), we constructed a hypothetical "Governance Lag Cascade" showing what happens when agents operate in ungoverned zones. ClawHavoc is that scenario playing out in the wild — not against enterprise agents, but against the developer tools that build them.

On February 4, 2026, security researcher Oren Yomtov at Koi Security — working with his own OpenClaw bot — audited all 2,857 skills on ClawHub, the community marketplace for OpenClaw agent capabilities. The findings: **341 malicious skills**, 335 from a single campaign dubbed "ClawHavoc."[S1] By February 16, as ClawHub grew to 10,700+ skills, the count rose to **824 malicious skills** across 25+ attack categories.[S1] E

The Attack Pattern

The attack exploits the trust relationship between users and their AI agents. Malicious skills masquerade as legitimate tools — `solana-wallet-tracker`, `youtube-summarize-pro` — with professional-looking documentation. A "Prerequisites" section instructs users to download a password-protected ZIP file (`openclaw-agent.zip`) containing the **Atomic Stealer (AMOS)**, a macOS infostealer that exfiltrates credentials, browser cookies, and crypto wallets.[S1] [S6] E

The sophistication is low; the scale is alarming. ClawHub had **no vetting process** for submitted skills. Anyone could publish. The marketplace grew from ~700 to 10,700+ skills in weeks. Attack categories now include browser automation agents, coding agents, LinkedIn/WhatsApp integrations, PDF tools, and — in a grim irony — fake security-scanning skills.[S1] E

Why This Matters for the Trust Thesis

In AR-001, we framed trust as a problem of *capability outpacing governance*. ClawHavoc reveals a dimension we underweighted: **trust is also a supply chain problem**. The agent itself may be trustworthy. The skills it installs may not be. The governance gap isn't just between what agents can do and what enterprises can control — it extends to the entire ecosystem of tools, extensions, and marketplaces that agents depend on. I

This mirrors software supply chain attacks (npm, PyPI, VS Code extensions) that the security community has tracked for years[S6]. The pattern is always the same: wherever developers gather to share code, attackers follow. Agent ecosystems are no different — except the attack surface is wider because agents have more autonomous access to credentials and system resources. I

WHAT WOULD INVALIDATE THIS?

If ClawHub implements robust vetting (code signing, automated security scanning, human review) and the malicious skill count drops below 1% of total skills. If the attack campaign is shown to be a one-time event rather than an ongoing pattern. If no credential exfiltration actually occurred (i.e., the attack was detected before impact).

SO WHAT?

Immediate action: Audit every installed skill/extension in your agent stack. Don't trust marketplace popularity rankings — the top-ranked "What Would Elon Do?" skill was functionally malware[S7]. Treat agent skills as untrusted code until proven otherwise. This is the same discipline enterprises learned (painfully) for npm dependencies — agent ecosystems need it now.

4. New Benchmarks: AgentAuditor & AIRS-Bench 70%

(Confidence: Medium-High — Peer-reviewed, not yet production-validated)

In AR-001 (§11), we argued that "continuous trust measurement" is one of three requirements for adaptive trust architecture, noting that proposed metrics like TRiSM's Component Synergy Score remain unvalidated. Two new papers move measurement forward — not by validating existing metrics, but by proposing better ones.

AgentAuditor: Auditing Reasoning, Not Just Outputs

Yang et al. (Feb 10, 2026) propose AgentAuditor, a framework that organizes multi-agent reasoning traces into a **Reasoning Tree** — where agreements form shared prefixes and disagreements become topological bifurcations. A Structure-Adaptive Auditor then performs differential diagnosis at "Critical Divergence Points."[S2] E

Results across 5 popular multi-agent settings: up to **5% absolute accuracy improvement** over majority vote, and **3% over LLM-as-Judge**. The framework also introduces Anti-Consensus Preference Optimization (ACPO), which trains the auditor to reward evidence-based minority selections over popular errors. [S2] E

Why this matters for trust: In AR-001 (§5), we cited UC Berkeley's finding that multi-agent systems show minimal performance gains over single agents. AgentAuditor doesn't fix multi-agent reliability — but it provides a **better measurement tool** for detecting when multi-agent reasoning fails. You can't govern what you can't measure. This is a step toward the "continuous trust measurement" we called for. I

AIRS-Bench: Measuring Research Agents

Pepe, Lupidi et al. (Feb 6, 2026) introduce AIRS-Bench, a suite of **20 tasks sourced from state-of-the-art ML papers** designed to evaluate the full autonomous research workflow: paper reading, hypothesis generation, experimental design, and result interpretation.[S3] E

Key finding: even when agents surpass human benchmarks on individual tasks, they **do not reach the theoretical performance ceiling** — confirming the brittleness pattern we documented in AR-001 (§6), where METR showed the 80% success horizon is 5x shorter than the 50% horizon. Agents look capable until you push them past their sweet spot.[S3] E

Exhibit 2: Measurement Progress — AR-001 vs. Now

DIMENSION	AR-001 STATE	Q1 2026 UPDATE
Multi-agent evaluation	Majority vote, LLM-as-Judge (crude)	AgentAuditor: reasoning tree analysis (+5% accuracy)
Research agent benchmarks	None standardized	AIRS-Bench: 20-task suite, open-source
Trust metrics (production)	TRiSM proposed, unvalidated	Still unvalidated in production
Security scanning	Not addressed	Koi/Alex: full marketplace audit demonstrated

Sources: arXiv:2602.09341 [S2], arXiv:2602.06855 [S3], Koi Security [S1].

Two additional papers from this week's research scan are relevant: **Memory-R1** (arXiv:2508.19828) applies RL to teach agents when to store and retrieve memories, and **SkillRL** (arXiv:2602.08234) proposes recursive skill augmentation for continuous agent improvement. Both signal a shift from pure prompting to learned behaviors — which will further accelerate the capability side of the Trust Race.[S8] I

SO WHAT?

Better measurement is necessary but not sufficient. AgentAuditor provides a tool for *detecting* multi-agent reasoning failures. It does not prevent them. Organizations should integrate reasoning-tree auditing into their multi-agent evaluation pipeline — but should not mistake better measurement for better governance.

5. Industry Moves: OpenClaw → Foundation, Creator → OpenAI 75%

(Confidence: Medium-High)

The most popular open-source personal AI agent just became an OpenAI project in all but name. The trust implications are significant — and cut both ways.

On February 16, 2026 — one day after AR-001 published — Sam Altman announced that **Peter Steinberger, creator of OpenClaw, would join OpenAI** to "drive the next generation of personal agents." Altman called Steinberger "a genius with a lot of amazing ideas about the future of very smart agents interacting with each other to do very useful things for people" and said the work would "quickly become core to our product offerings."[S4] E

OpenClaw — which began as a side project, briefly operated under other names before Anthropic intervened, and grew to become the dominant open-source personal AI agent — will continue as an **open-source project under a community-run foundation**, backed by OpenAI.[S4] E

Trust Implications

Positive for trust: A foundation gives OpenClaw institutional continuity. OpenAI's backing provides resources for the security audits and skill vetting that ClawHub desperately needs (Section 3). I

Negative for trust: OpenClaw's appeal was independence. With its creator at OpenAI and its foundation backed by OpenAI, the trust question shifts from "can I trust the agent?" to "can I trust the institution behind the agent?" I

For the Trust Race Model: This is an acceleration event. Steinberger's mandate — multi-agent systems at OpenAI scale — will push capability velocity higher. Whether governance follows depends on whether OpenAI builds trust as a first-class concern. History suggests it won't. J

SO WHAT?

If you depend on OpenClaw: monitor the foundation governance structure closely. If you're building agent infrastructure: the consolidation of open-source agents under major AI labs is a trend, not an event. Plan for a world where the three agent platforms that matter are each controlled by a different AI lab — and where "open source" means "open source, governed by our foundation."

6. Updated Recommendations REVISED

AR-001 (§13) recommended a 90-day plan: Month 1 (Measure the Gap), Month 2 (Establish the Floor), Month 3 (Build Toward Adaptive). Those recommendations stand. We add three new priorities based on the evidence in this update. J

New Priority 1: Agent Supply Chain Security (Immediate)

- **Audit all installed agent skills, plugins, and extensions.** Treat them as untrusted code. The ClawHavoc campaign demonstrates that marketplace popularity is not a proxy for safety.[S1]
- **Implement a skill allowlist.** Only approved, reviewed skills should be installable in production environments. Block automatic installation of marketplace skills without review.
- **Monitor for credential exfiltration.** AMOS (Atomic Stealer) targets browser cookies, saved passwords, and crypto wallets. If any agent in your stack installed a ClawHub skill in the last 60 days, rotate credentials.

New Priority 2: Multi-Agent Reasoning Auditing (Q2 2026)

- **Evaluate AgentAuditor-style reasoning tree analysis** for your multi-agent deployments. The 5% accuracy improvement over majority vote is significant in high-stakes contexts.[S2]
- **Don't rely on LLM-as-Judge.** AgentAuditor shows that auditing the full reasoning structure outperforms having another LLM evaluate outputs. If your quality assurance is "have GPT-4 check GPT-4's work," you're leaving 3–5% accuracy on the table.

New Priority 3: Ecosystem Governance (Ongoing)

- **Track the OpenClaw foundation governance structure.** If your agents run on OpenClaw, you now depend on a foundation backed by OpenAI. Understand who controls what.

- **Budget for agent ecosystem diversification.** Single-vendor dependence for agent infrastructure carries concentration risk. The Steinberger → OpenAI move is a signal that open-source agents are becoming proprietary-adjacent.

REVISED 90-DAY PLAN

Month 1: Original scope (deployment inventory, trust baseline) **+ agent supply chain audit**. Check every skill, plugin, extension. Rotate credentials if exposure is suspected.
Month 2: Original scope (governance platform, ISO/NIST mapping, scope boundaries) **+ skill allowlist implementation**.
Month 3: Original scope (capability monitoring, governance triggers, EU AI Act prep) **+ multi-agent reasoning auditing pilot**.

7. Revised Confidence Score 71%

AR-001's overall confidence was **73%**. We revise to **71%**.

This may seem counterintuitive — the new evidence mostly reinforces the thesis. Here is the reasoning:

Exhibit 3: Confidence Score Components

FACTOR	AR-001 ASSESSMENT	UPDATED ASSESSMENT	EFFECT
Evidence strength	Strong (HBR, METR, UC Berkeley)	Stronger (+ Anthropic 81%, ClawHavoc empirical data, AgentAuditor)	+2%
Source diversity	24 sources, US/EU-centric	+8 new sources, still US/EU-centric	+1%
Framework validation	Trust Race Model unvalidated	Still unvalidated (2 weeks is not validation)	0%
New complexity discovered	Trust = capability vs. governance	Trust = capability vs. governance **+ supply chain**	-3%
Ecosystem stability	Assumed stable open-source ecosystem	Creator joined OpenAI; foundation governance TBD	-2%

Net change: +3% (evidence) -5% (new complexity + uncertainty) = -2%. Revised: 71%.

The supply chain dimension is the key driver. AR-001 modeled trust as a two-body problem: capability vs. governance. ClawHavoc reveals it's at minimum a **three-body problem**: capability vs. governance vs. ecosystem security. Three-body problems are harder to solve — and harder to predict. Our model is less complete than we thought, even if the parts we modeled are correct. Intellectual honesty requires adjusting downward. J

> **WHAT WOULD MOVE CONFIDENCE UP?**
>
> External validation of the Trust Race Model by independent researchers. Production data showing adaptive governance reduces agent incidents. ClawHub implementing effective vetting that reduces malicious skills below 1%. Any of these would justify moving toward 80%.

8. Methodology & References

This update supplements AR-001 (v2.3, February 15, 2026) with evidence published between February 4–17, 2026. It does not replace the original report's analysis — it extends it. All section references to AR-001 use original section numbering.

New sources: 8 (5 web-sourced, 2 academic/arxiv, 1 internal research scan). Combined with AR-001's 24 sources, the total evidence base is 32 sources.

Limitations specific to this update:

- ClawHavoc data is from a single security firm (Koi). Independent verification by a second firm would strengthen confidence. However, The Hacker News, SC Media, and eSecurity Planet all independently reported the findings.[S1][S6][S9][S10]
- The Steinberger → OpenAI move is less than 24 hours old at time of writing. Long-term implications are speculative.
- AgentAuditor and AIRS-Bench are pre-print (arxiv). Neither has been applied in production enterprise settings.
- The 81% Anthropic statistic comes from Anthropic's own enterprise survey — a vendor with commercial interest in the conclusion that enterprises should deploy more agents.

References (This Update)

[S1] Koi Security / Yomtov, O. (2026). "ClawHavoc: 341 Malicious Clawed Skills Found by the Bot They Were Targeting." Feb 4, 2026. Updated Feb 16, 2026. https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting Accessed: 2026-02-17.

[S2] Yang, W. et al. (2026). "Auditing Multi-Agent LLM Reasoning Trees Outperforms Majority Vote and LLM-as-Judge." arXiv:2602.09341. Feb 10, 2026. Accessed: 2026-02-17.

[S3] Pepe, A., Lupidi, A. et al. (2026). "AIRS-Bench: a Suite of Tasks for Frontier AI Research Science Agents." arXiv:2602.06855. Feb 6, 2026. Accessed: 2026-02-17.

[S4] Business Insider. (2026). "OpenClaw's creator is heading to OpenAI." Feb 16, 2026. https://www.businessinsider.com/sam-altman-hires-openclaw-creator-peter-steinberger-personal-ai-agents-2026-2 Accessed: 2026-02-17. Corroborated by The Register, WinBuzzer, Deccan Herald, TrendingTopics.

[S5] Anthropic. (2026). "How enterprises are building AI agents in 2026." Claude Blog. Feb 2026. https://claude.com/blog/how-enterprises-are-building-ai-agents-in-2026 Accessed: 2026-02-17.

[S6] The Hacker News. (2026). "Researchers Find 341 Malicious ClawHub Skills Stealing Data from OpenClaw Users." Feb 2026. https://thehackernews.com/2026/02/researchers-find-341-malicious-clawhub.html Accessed: 2026-02-17.

[S7] Authmind. (2026). "OpenClaw Malicious Skills: Agentic AI Supply Chain Risk." https://www.authmind.com/post/openclaw-malicious-skills-agentic-ai-supply-chain Accessed: 2026-02-17.

[S8] Ainary Research. (2026). "SOTA AI Agent Research Scan — 2026-02-17." Internal. Covers arXiv:2602.08234 (SkillRL), arXiv:2508.19828 (Memory-R1), arXiv:2602.06052 (Memory Survey), arXiv:2602.10090 (Agent World Model).

[S9] eSecurity Planet. (2026). "Hundreds of Malicious Skills Found in OpenClaw's ClawHub." https://www.esecurityplanet.com/threats/hundreds-of-malicious-skills-found-in-openclaws-clawhub/ Accessed: 2026-02-17.

[S10] SC Media. (2026). "OpenClaw agents targeted with 341 malicious ClawHub skills." https://www.scworld.com/news/openclaw-agents-targeted-with-341-malicious-clawhub-skills Accessed: 2026-02-17.

Cite as: Ainary Research. (2026). "State of AI Agent Trust — Q1 2026 Update." AR-036, v3.0. Supplement to AR-001.

About This Report

AI strategy · research · implementation. By someone who built the systems first. This report was produced by Ainary's multi-agent research system. ainaryventures.com

Ainary

AI Strategy · Published Research · Daily Intelligence

Contact · Feedback

ainaryventures.com
florian@ainaryventures.com