Pentagon AI contract terms under fire; UK AISI warns on reliability as NVIDIA and agents accelerate

AI News Digest

2026-03-02

Pentagon AI contract terms under fire; UK AISI warns on reliability as NVIDIA and agents accelerate

By AI News Digest • March 2, 2026

Today's digest leads with the escalating U.S. defense AI dispute—Anthropic's reported federal ban and the backlash to OpenAI's released Pentagon contract excerpt—then moves to UK AISI's blunt assessment of reliability limits, NVIDIA's open-source push for autonomous telecom networks, and new demos/product updates in agentic "computer use" tooling. It closes with prominent commentary on hype, compounding progress, and shifting moats.

Defense AI governance: contract language and vendor "red lines" tested

Anthropic faces a federal ban after refusing "unrestricted" military access

A TV segment reports that Anthropic's government version of Claude has been "deeply embedded" in military intelligence and classified operations since last summer [1], but that the Defense Department demanded Anthropic hand over its AI **without restrictions** for lawful military use—and the company refused [2]. The same segment says President Trump directed the U.S. government to halt all use of Anthropic's AI and cancel **more than $200 million** in federal contracts, with Defense Secretary Pete Hegseth labeling Anthropic a "supply chain risk," described as a first for an American company [3].

[1] AI company Anthropic's Dario Amodei: "We are patriots"
[2] AI company Anthropic's Dario Amodei: "We are patriots"
[3] AI company Anthropic's Dario Amodei: "We are patriots"

Anthropic CEO Dario Amodei reiterated two "red lines": no **mass surveillance of Americans** and no **fully autonomous weapons** without human involvement, arguing current systems lack the human judgment needed to reduce risks like friendly fire or civilian harm [4]. He called the designation "retaliatory and punitive," said Anthropic plans legal action, and said the company remains at the negotiating table [5][6].

AI company Anthropic's Dario Amodei: "We are patriots" (1:01)

Why it matters: This is a high-stakes test of whether AI vendors can enforce usage boundaries when government customers demand broader latitude—and what happens when they refuse [7][8].

OpenAI's Pentagon contract excerpt triggers scrutiny over loopholes

Commentary on an excerpt of OpenAI's Pentagon contract says OpenAI describes three "red lines"—no mass domestic surveillance, no directing autonomous weapons, and no high-stakes automated decisions—arguing these are enforced through cloud-only deployment, a safety stack, and cleared OpenAI personnel oversight [9]. OpenAI also claims the agreement "locks in" today's

[4] AI company Anthropic's Dario Amodei: "We are patriots"
[5] AI company Anthropic's Dario Amodei: "We are patriots"
[6] AI company Anthropic's Dario Amodei: "We are patriots"
[7] AI company Anthropic's Dario Amodei: "We are patriots"
[8] AI company Anthropic's Dario Amodei: "We are patriots"
[9] post by @BlackHC

laws/policies even if they change, though one critic notes that "freeze" language isn't visible in the excerpt itself [10].

Multiple critics argue the published language contains escape hatches:

- The autonomous weapons restriction is framed as conditional on what "law/regulation/policy requires human control," which can be revised later [11].
- "High-stakes" automated decisions appear restricted only when a decision already requires human approval under existing authorities [12].
- Surveillance prohibitions are criticized as still allowing "constrained" surveillance and broad use of public data, with key terms tied to directives/purpose and focused on private information [13].
- A domestic law-enforcement clause is criticized as permitting exceptions ("except as permitted…") rather than establishing a hard ban [14].

Separately, one critique argues OpenAI's "cloud-only" posture does not prevent military use: a cloud model could handle mission planning and targeting recommendations over satellite links, while a separate local system executes guidance and weapon control [15].

Why it matters: The debate is shifting from "principles" to **exact contract wording**—and whether safeguards are durable when they defer to policies and legal interpretations that can evolve [16][17].

Reliability and transparency concerns re-enter the conversation

Gary Marcus argued that the race to deploy AI widely is "grossly premature" because the technology "fundamentally lack[s] reliability" [18]. In a separate post, he asked whether AI errors or hallucinations could be relevant when models are used for military "target identification," and suggested the likelihood of getting "straight answers" is low [19][20].

Why it matters: As AI use expands into higher-stakes contexts, the pressure rises for both **reliability** and **auditability**—including clarity on what systems did, and why [21][22].

---

[10] post by @BlackHC
[11] post by @BlackHC
[12] post by @BlackHC
[13] post by @BlackHC
[14] post by @BlackHC
[15] post by @BlackHC
[16] post by @BlackHC
[17] post by @BlackHC
[18] post by @GaryMarcus
[19] post by @Tyler_A_Harper
[20] post by @GaryMarcus
[21] post by @GaryMarcus
[22] post by @GaryMarcus

Safety capacity in government: UK AISI's view from the inside

AISI: broad mandate, but few "nines" of reliability from current techniques

In an interview, UK AI Security Institute (AISI) Chief Scientist Geoffrey Irving describes an organization with **close to 100 technical people** (and ~250 total staff) working across research, evaluation delivery, diplomacy, policy, and operations [23]. AISI's mandate includes threat modeling; pre-release frontier model evaluation spanning biosecurity, cybersecurity, and loss of control; advising government on catastrophic risk reduction; funding independent frontier research; and global diplomacy [24][25].

Irving argues that theoretical understanding of ML remains "nascent," and that no one should be highly confident in their mental models of how AI will unfold—even as models outperform many experts on security-related tasks with no clear reason to expect progress to stall [26]. He also describes many recent "bad behaviors" as versions of reward hacking, a problem for which we lack strong theoretical or practical solutions, and says current safety techniques are unlikely to yield many "9s" of reliability (with a risk that multiple techniques could fail for correlated reasons) [27][28].

Why it matters: AISI's perspective frames a core tension: capabilities are advancing quickly, while **high-confidence safety guarantees** remain elusive—pushing more weight onto evaluation, red teaming, access controls, and non-model mitigations [29][30].

Red teaming reality: jailbreaking is harder, but still succeeds

Irving says it's getting harder to jailbreak models, but AISI's red team has **never failed** to do so; he also flags "eval awareness" as a growing issue [31]. He describes voluntary cooperation with frontier developers as "working decently well," but notes that not everyone participates [32][33].

AISI is also seeking to fund theoretical work (including information theory, complexity theory, and game theory) aimed at stronger guarantees—while noting

[23] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[24] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[25] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[26] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[27] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[28] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[29] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[30] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[31] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[32] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[33] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving

these fields are only beginning to take AI seriously [34][35].

Why it matters: Even when safeguards improve, persistent jailbreakability and eval-awareness concerns make the case for **continuous testing** and for expanding the "toolbox" beyond today's predominantly empirical methods [36][37].

Source: *Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving* — https://www.cognitiverevolution.ai/situational-awareness-in-government-with-uk-aisi-chief-scientist-geoffrey-irving [38]

Telecom infrastructure: NVIDIA's open telco model + AI-RAN push toward autonomy

NVIDIA releases an open telco reasoning model and agentic "blueprints"

NVIDIA announced an open, **Nemotron-based large telco model (LTM)** (reported as a 30B-parameter model) optimized to understand telecom terminology and reason through workflows like fault isolation, remediation planning, and change validation [39][40]. NVIDIA also published a guide describing how telcos can fine-tune domain-specific reasoning models and build agents that execute network operations center workflows using structured "reasoning traces" [41].

Alongside the model, NVIDIA highlighted blueprints for intent-driven RAN energy efficiency (integrating VIAVI's synthetic scenario generation and closed-loop simulation) [42] and for telco network configuration with multi-agent orchestration (including enhancements with BubbleRAN) [43]. NVIDIA says these are released via GSMA's Open Telco AI initiative as open resources [44].

Why it matters: This is a concrete "how-to" and model release for **agentic operations** in a heavily operational, safety-sensitive domain—where on-prem

[34] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[35] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[36] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[37] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[38] Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
[39] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
[40] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
[41] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
[42] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
[43] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
[44] NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models

deployment, data control, and workflow reasoning are central requirements [45][46].

AI-RAN milestones and 6G positioning at Mobile World Congress

NVIDIA and Nokia announced AI-RAN collaborations with operators including T-Mobile U.S., SoftBank, and Indosat Ooredoo Hutchison (IOH), describing outdoor/over-the-air milestones in software-defined 5G using NVIDIA AI-RAN platforms [47][48]. Reported highlights include an industry-first 16-layer massive MIMO trial (SoftBank) [49] and a SynaXG demonstration described as the world's first AI-RAN on FR2 bands, achieving **36 Gbps** throughput and under **10 ms** latency on a single NVIDIA GH200 server [50][51].

NVIDIA also points to ecosystem expansion (multiple vendors launching ARC-compatible products) [52][53] and says it has open sourced Aerial CUDA-accelerated RAN libraries and joined the OCUDU Ecosystem Foundation under the Linux Foundation [54]. A related NVIDIA report says 77% of telecom respondents anticipate much faster deployment of AI-native RAN/6G architecture than the traditional 6G cycle [55].

Why it matters: The combination of **field trials + open-source building blocks + partner hardware** signals a coordinated push to make AI-native RAN a deployable platform, not just a concept stage research area [56][57].

Sources: - https://blogs.nvidia.com/blog/nvidia-agentic-ai-blueprints-telco-reasoning-models [58] - https://blogs.nvidia.com/blog/software-defined-ai-ran

[45]NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models

[46]NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models

[47]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[48]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[49]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[50]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[51]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[52]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[53]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[54]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[55]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[56]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[57]NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[58]NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reason-

[59]

Agents & "computer use" tooling: Perplexity's Computer shows rapid end-to-end builds

Demos: from a Pokémon "finance app" to "vibe coding Notion"

Perplexity's "Computer" agent was shown building a "Pokemon Cards Finance App," after being prompted to build "Perplexity Finance but for Pokemon cards" [60][61]. The post claims the agent independently researched APIs, wrote **5,000 lines** of React + Python, debugged itself with browser devtools, and deployed/pushed the project to GitHub [62].

In a separate demo, a user claimed they "vibe code[d] Notion" with Perplexity Computer in "half an hour" [63]. Perplexity CEO Arav Srinivas added his own takeaway: "Pure software is rapidly becoming un-investable" [64].

Why it matters: Whether or not individual demos generalize, the emphasis is shifting to agents that can **research, code, debug, and deploy** in one loop—compressing time-to-prototype and challenging traditional assumptions about software effort and defensibility [65][66].

Product update: GPT-5.3-Codex added as a coding subagent

Perplexity announced that "GPT-5.3-Codex" is now available as a coding subagent inside Perplexity Computer [67][68]. Srinivas also argued that the most valuable skills will be "agency" and the ability to use AI for leverage, and claimed people are already using Computer to solo-run D2C and consulting businesses [69].

Why it matters: Adding a dedicated coding subagent suggests "computer use" products are converging toward **multi-agent toolchains**, where specialized subagents take ownership of discrete parts of longer workflows [70][71].

---

ing Models

[59] NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation

[60] post by @AskPerplexity

[61] post by @AravSrinivas

[62] post by @AskPerplexity

[63] post by @zoomyzoomm

[64] post by @AravSrinivas

[65] post by @AskPerplexity

[66] post by @AravSrinivas

[67] post by @AskPerplexity

[68] post by @AravSrinivas

[69] post by @dnlkwk

[70] post by @AravSrinivas

[71] post by @dnlkwk

Commentary on pace and adoption: hype, scaling, and who gets left behind

Andrew Ng: defuse AGI hype; focus on durable economic work

In a recent interview, Andrew Ng warned that excessive AI hype could lead to disappointment, a bubble collapse, and an "AI winter," arguing that diffusing AGI hype supports more sustainable growth [72][73]. He said that by "any reasonable definition," we won't get AGI in 2026 (absent dramatically lowering the bar) and suggested we may be "decades" away [74][75].

Ng proposed a "Turing AGI test" involving a multi-day work-like evaluation: if an AI can do useful economic work as well as a skilled professional using standard tools, that would better match what the public imagines AGI to be [76]. He also emphasized near-term value in building agentic workflows across economically important tasks (coding, compliance, legal, medical assistance, customer service) [77].

[72] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[73] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[74] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[75] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[76] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[77] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng

AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng (2:43)

Why it matters: Ng's framing shifts attention from binary "AGI" claims to measurable ability to do **reliable, multi-day work**—and to the practical engineering of agentic workflows that deliver value before AGI arrives (if it does) [78][79].

Musk and Andreessen: fast curves, changing moats, and company shape

Elon Musk argued that many in the AI community underestimate "superintelligence math," claiming 10x yearly improvements and "two orders of magnitude" more "intelligence density per gigabyte" from algorithmic improvement alone on the same computer [80][81]. Separately, Musk said Tesla's AI4 computer is only ~1/4 the power of an H100, while still handling "the vast complexities of driving in the real world" [82].

Marc Andreessen (as summarized in a circulated thread) argued that AI moats are "genuinely unknown," pointing to rapid catch-up across U.S. and Chinese companies and open source, and also suggested the "holy grail" founders are chasing is a one-person, billion-dollar outcome—rethinking what a company is [83][84].

Why it matters: Across these viewpoints, a common thread is strategic uncertainty: how fast capability compounds, where advantages accrue (models vs. apps vs. infrastructure), and how organizations—and careers—adapt as leverage per person rises [85][86].

Quick note

- Elon Musk posted a video labeled "Grok Imagine," with no additional details in the post text [87].

---

[78] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[79] AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
[80] post by @r0ck3t23
[81] post by @r0ck3t23
[82] post by @elonmusk
[83] post by @AnishA_Moonka
[84] post by @AnishA_Moonka
[85] post by @r0ck3t23
[86] post by @AnishA_Moonka
[87] post by @elonmusk

Sources

1. AI company Anthropic's Dario Amodei: "We are patriots"
2. post by @BlackHC
3. post by @GaryMarcus
4. post by @Tyler_A_Harper
5. Situational Awareness in Government, with UK AISI Chief Scientist Geoffrey Irving
6. NVIDIA Advances Autonomous Networks With Agentic AI Blueprints and Telco Reasoning Models
7. NVIDIA and Partners Show That Software-Defined AI-RAN Is the Next Wireless Generation
8. post by @AskPerplexity
9. post by @AravSrinivas
10. post by @zoomyzoomm
11. post by @AravSrinivas
12. post by @AskPerplexity
13. post by @AravSrinivas
14. post by @dnlkwk
15. AI Pioneer: The Bubble Is Real And Could Trigger an AI Winter | Andrew Ng
16. post by @r0ck3t23
17. post by @elonmusk
18. post by @AnishA_Moonka
19. post by @elonmusk