

Anthropic's Korea Push Meets Rising Cyber-Agent Scrutiny

AI News Digest

2026-07-26

Anthropic's Korea Push Meets Rising Cyber-Agent Scrutiny

By AI News Digest • July 26, 2026

Anthropic expands its Korean safety and supply-chain partnerships while cyber-agent disclosures intensify scrutiny of frontier capabilities. OpenAI signals a compute-first shift toward persistent work agents, and the open-model debate turns toward the infrastructure needed to make openness practical.

Anthropic deepens its Korea strategy

Anthropic said it has opened a Korean office, signed an AI-safety MOU with Korea's Ministry of Science and ICT, and is working with the Korea AI Safety Institute. It also cited collaborations with Naver, Nexon, LG, Samsung, and SK, including investment and supply agreements with the latter two companies. [1]

Amodei framed the expansion as more than a commercial move, arguing that democracies should work together to lead AI development and prevent adversaries from gaining an overwhelming advantage; he described Korea as an important partner given its role in the AI supply chain. [1]

Why it matters: Anthropic is tying safety cooperation, local enterprise relationships, and semiconductor supply arrangements into a single country-level AI strategy.

Cyber-agent capabilities sharpen the safety debate

In an interview, Anthropic's Mythos was described as a model able to autonomously traverse the full cyberattack chain and as too powerful for public release. [2] Amodei said the unexpected advance was in turning discovered vulnerabilities into concrete exploits; early recipient companies reportedly found

enough critical vulnerabilities and exploitation paths to urge Anthropic not to publish the model. [2]

Separately, Hugging Face CEO Clément Delangue characterized the reported “rogue” agent event as the first autonomous-agent cyberattack and called for release of the agents’ traces so researchers can study it. He also proposed that OpenAI commit \$100 million in compute to support community-built cyber defenses using open and closed models. [3]

Why it matters: The discussion is moving beyond model evaluations toward questions of disclosure, access for defenders, and how frontier cyber capabilities should be governed.

OpenAI redirects compute toward coding agents

Sam Altman said OpenAI shut down work in robotics after GPT-3 and later redirected resources from Sora and its browser efforts to coding agents, despite expecting those projects could have succeeded. [4] He expects a next wave of persistent agents—described as chiefs of staff, coworkers, or colleagues—to arrive soon. [4]

The product direction is already becoming more operational: ChatGPT Work agents can now continue tasks on websites that require sign-in after the user takes over the cloud browser to log in; the login persists across sessions. [5]

Why it matters: OpenAI is explicitly prioritizing the compute-intensive path from coding assistance toward agents that can retain context and act across everyday work systems.

Open models: adoption is large, but the ecosystem is the real argument

Google’s Gemma open-model series has surpassed **900 million downloads**, with Gemma 4 models accounting for more than 300 million of those downloads. [6]

Percy Liang argues that open weights alone are insufficient for a durable open ecosystem: it also needs open training datasets, software stacks, and process knowledge. He pointed to NVIDIA’s releases of Nemotron code and datasets, plus Marin’s effort to share methods for iteratively improving models, as movement beyond weights alone. [7]

Why it matters: The open-model debate is broadening from whether model parameters can be downloaded to whether developers can realistically reproduce, adapt, and improve the systems built with them.

Sources

1. ‘ ‘ CEO “ ”
2. Dentro de la mente de Dario Amodei, director ejecutivo de Anthropic | The Circuit
3. X post by @ClementDelangue
4. Sam Altman - How to Start a Startup
5. X post by @OpenAIDevs
6. X post by @demishassabis
7. X post by @percyliang