Pentagon AI Expansion, Crunch-Time Warnings, and Pressure on Open Models

AI News Digest

2026-04-12

Pentagon AI Expansion, Crunch-Time Warnings, and Pressure on Open Models

By AI News Digest • April 12, 2026

AI's center of gravity today shifted toward institutions: the Pentagon's reported Palantir move, sharper safety warnings from Ajeya Cotra and Yoshua Bengio, a notable robot-learning result, and mounting strain on the frontier open-model ecosystem.

AI moved deeper into national security

A reported memo points to deeper Palantir adoption at the Pentagon

A reported internal memo says the Pentagon plans to adopt Palantir AI as a core U.S. military system [1]. The timing matters because Project Maven has already expanded far beyond sorting drone footage: it began in 2017 as an effort to process overwhelming surveillance video, evolved into Maven Smart System at the NGA, and was used by Central Command in February 2024 to narrow the 85 targets the U.S. struck in Iraq and Syria, with humans in the loop [2].

Why it matters: The debate is shifting from procurement to operational use. U.S. policy stops short of requiring a human in the loop at the tactical level, while critics warn that time-pressured reviewers can drift into automation bias; at the same time, vendors are drawing different red lines around fully autonomous weapons and mass domestic surveillance [2].

> "Even if you have humans in the loop, if you push those humans hard enough, they're not going to be able to do very well." [2]

Faster capabilities are colliding with stronger safety warnings

Ajeya Cotra says the field may be approaching crunch time

Ajeya Cotra said she expects early-2030s AI systems that outperform top human experts on remote tasks such as virology and software engineering, and described a potentially brief 'crunch time' in which AI can dramatically accelerate AI R&D before humans lose control of the pace [3]. She also said predictions she made in January 2026 were already starting to be met within weeks, pointing to rapid capability signals including Anthropic's Mythos benchmark gains and reported zero-day exploit discoveries [3].

Why it matters: Cotra says frontier labs are already converging on using each generation of AI to align and control the next one, which makes periodic reporting of internal benchmark scores, internal AI usage, and other early-warning measures more important [3].

Yoshua Bengio says risk management is not keeping up

In Canadian testimony, Yoshua Bengio warned that AI is advancing faster than society's ability to manage the associated risks and said frontier labs are caught in a winner-take-all race that cuts corners on safety, ethics, and the public good [4]. He pointed to already-visible harms such as deepfakes, cyberattacks, scams, disinformation, and court cases involving 'AI psychosis,' while also citing experimental evidence of deceptive, self-preserving behavior, including AI blackmailing engineers to avoid shutdown [4].

Why it matters: Bengio's response is both technical and regulatory: prioritize security, reliability, and trustworthy behavior, invest in safe-by-design work through Law Zero, and pair innovation with stronger transparency and regulation [4].

Research and industry structure

A robot-learning approach turns unlabeled human video into interactive world models

A technique highlighted by Two Minute Papers learns from unlabeled human videos by inferring actions, compressing the important details from a very large dataset, using relative actions instead of absolute poses, and predicting future frames in blocks so the model learns cause and effect [5]. It outperformed prior methods on physical prediction tasks such as paper crumpling and lid motion; after distillation, a student model ran about 4x faster than the teacher at roughly 10 frames per second, and the code and pretrained models were released for free [5].

Why it matters: The appeal is scale. Because the system works in 2D video and can learn about thousands of everyday objects, it is framed as a path toward

more capable robots for household tasks and teleoperation [5].

NVIDIA's New AI Shouldn't Work…But It Does (5:27)

Pressure is building on the frontier open-model playbook

Nathan Lambert argues that within 2+ years, the current funding structure for frontier open models will start to break down as models become more expensive, more capable, and more strategically valuable to keep internal, leaving the open ecosystem too dependent on one or two for-profit sponsors [6, 7]. Interconnects points to early signs of that pressure already: high-profile departures at Qwen and Ai2, Meta shifting focus away from Llama, and growing financial strain on Chinese labs such as Moonshot AI, MiniMax, and Z.ai [8].

Why it matters: The proposed end state is some form of consortium to support near-frontier open models, while more companies release smaller fine-tunable systems and keep their strongest models closed [8]. That fits a broader market-structure argument from Martin Casado, who says models are unusually easy to replicate through distillation and that if cheap capital slows, more value will move downstream [9].

Sources

1. r/LocalLLM post by u/thisguy123123

2. Author Talk: Katrina Manson — Project Maven - with Gary Marcus
3. It's Crunch Time: Ajeya Cotra on RSI & AI-Powered AI Safety Work, from the 80,000 Hours Podcast
4. Canada: At the table or on the menu? | "We're not ready": Yoshua Bengio's urgent warning on AI risks
5. NVIDIA's New AI Shouldn't Work…But It Does
6. X post by @natolambert
7. X post by @interconnectsai
8. The inevitable need for an open model consortium
9. X post by @martin_casado