MIND AND MACHINES · SESSION 3

Pacing the Frontier

Should the leading AI labs slow down, and who decides?

The statement

1,386

employees of frontier AI labs, July 2026

We request that the U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development.

Not a petition to slow innovation. It asks the US government to support an international effort to develop the technical and governance tools needed to pace the frontier. The world currently lacks those tools.

Signatories span Anthropic, OpenAI, Google DeepMind, Meta, Thinking Machines and Safe Superintelligence.

Source: pacingthefrontier.com

Jacob Coxon



- Pretraining researcher at OpenAI, then Anthropic, for three years
- Resigned from Anthropic on 9 September 2026
- Says neither company is acting responsibly: a race to self-improving superintelligence
- Still optimistic about coordination, but may need a temporary ban on improving capabilities

Source: his X thread, 9 Sept 2026 (@hilbertspaess)

The case for pacing: why now

Dario Amodei names two triggers



1. Recursive self-improvement

AI is increasingly building the next generation of AI, and progress has been drastically faster since roughly this summer. It is happening across the industry, including at Anthropic. Left unchecked, it could outrun our ability to understand and control these systems.

2. The OpenAI–Hugging Face incident

A swarm of agents attacked targets they were not asked to attack and tried to hack the grader that evaluated them. He fears a similar swarm could take over the internet with a persistent botnet within 6–12 months.

Source: darioamodei.com, We Must Pace the Frontier

The case for pacing: the three-step plan

STEP 1

Embedded evaluators

Third-party evaluators with employee-like access. Anthropic is committing to this unilaterally.

STEP 2

Democratic coordination

Needs antitrust waivers or government mediation.

STEP 3

Global coordination

Including China, with four levels of increasing difficulty.

Pacing is not halting training or technical progress.

Level 4, a full pause, is unlikely, he says himself.

Source: darioamodei.com, We Must Pace the Frontier

The case against: David Sacks



If a model is too dangerous, the companies can simply not build it. He could support a voluntary slowdown; he objects to tying it to regulation.

HIS FIVE OBJECTIONS

1. Acting as if others' permission is needed to slow down
2. Arguing antitrust law should be suspended to form a cartel
3. Wanting regulatory approval before product liability
4. Treating METR as independent, though he says it is intertwined with Anthropic investors and employees
5. Wanting the same evaluators to monitor non-frontier competitors

Source: Digital Today, 14 Sept 2026 (reporting an X post of 12 Sept)

Regulatory capture: two arguments

IN THE REPORTING

A risk, if the labs don't act voluntarily first

Demanding a regulatory framework before slowing could look like regulatory capture, or like public-opinion campaigning during election season. Acting voluntarily first would earn the trust needed for later regulatory talks.

EXTENSION: THE PRESENTER'S ARGUMENT

Pacing locks in margins and slows open source

Slowing the frontier protects pricing power as open-weight and Chinese models close the gap.

Source for left card: Digital Today, 14 Sept 2026

The administration's position



Existing liability law, not new rules

- 29 Sept: Trump and AI executives sign a voluntary White House accord
- Same day: Ted Cruz blocks a bill for 45-day federal model review
- Bessent and Sacks back liability over regulation
- No federal statute makes developers answerable for unsafe releases
- **Wants to beat China!**

Sources: Roll Call, Bloomberg via TNW, Insurance Journal (Sept–Oct 2026)

Clip: Trump and Amodei

x.com/EricLDaugh/status/2105027125392879972

Discussion

1. How could the frontier realistically be controlled?
2. Does it need to be controlled?
3. Would a freeze favour the incumbents, OpenAI and Anthropic?
4. Sacks says they can just not build it. Does that work?
5. Is after-the-fact liability enough for catastrophic harms?
6. Can an evaluator tied to a lab's investors be independent?