A Moment to Talk About AI

Why frontier AI labs need binding rules — and why the generation that understands these systems must help write them

Pierre Lague
PhD Student

September 2026

A year ago I wrote a proposal for a World AI Regulation Office (WAIRO): a message in a bottle addressed to anyone with the leverage to build a binding global authority for artificial intelligence. I expected it to be read, if at all, as a student's utopia. This week, the people building the most capable AI systems on Earth started saying roughly the same thing.

1 What the Insiders Say

On 8 September 2026, Jacob Coxon, a researcher who had spent three years working on model pre-training at OpenAI and then Anthropic, resigned publicly, writing that both companies are racing toward self-improving superintelligence and "gambling with our lives" [1, 2]. Public resignations like this are not new in the industry. What is new is that people who stayed agreed with him. Evan Hubinger, Anthropic's alignment science lead, answered: "we really do earnestly believe AI could kill all humans!" He added that he personally puts that probability above 10 percent within the next decade, and that the company does not yet have a plan to align superintelligence [2, 3]. Samuel Marks, who leads Anthropic's Cognitive Oversight team, wrote that "the more senior the employee, the more concerned they are", and that current techniques can only nudge models toward better behaviour rather than reliably align them [2, 4]. At OpenAI, a member of the safety team wrote publicly that she, too, thinks the field needs to slow down [5].

It is tempting to read this as proof that these companies simply do not care about human life. I think that reading is wrong — and, more importantly, less useful than the truth. Coxon's own diagnosis was structural. At one company, he said, staff have not internalised the stakes; at the other, staff understand them well but are locked in a race, each convinced that a rival would be less careful [2]. Marks pointed to the same forces: commercial pressure and fear of less responsible competitors [4].

That is a textbook collective-action problem. No lab can slow down alone without ceding ground, so none does. And collective-action problems have one reliable solution: a rule that binds every player at the same time, enforced by someone who is not a player.

The labs know this. On 28 July 2026, more than 1,200 employees across Anthropic, Google DeepMind, OpenAI and Meta — including Anthropic's CEO and OpenAI's chief scientist — published a statement asking the U.S. government to support an international effort to build the technical and governance tools needed to deliberately pace frontier AI development, tools the world does not currently have [6, 7]. Notably, the statement was organised with support from Encode, a youth-led advocacy group I return to below [6]. OpenAI's chief scientist has since written that he favours binding rules that would require every company to move at a more considered pace [2].

Meanwhile, the race continues. OpenAI told reporters that on 28 August it began training a model significantly more powerful than anything it has released, and both companies are preparing stock-market listings [2]. In early September, the reveal of OpenAI's Astra model

and its staggering 99.8% score on ARC-AGI 3 has shook a lot of people from the community. When the builders of a technology ask for brakes while still pressing the accelerator, they are not necessarily being hypocritical. They are telling us, as clearly as a competitor in a race can, that the brakes have to be installed from outside.

2 The risks are no longer hypothetical

For years, the standard objection to frontier regulation was that the risks were speculative. That objection gets weaker every month.

In July, OpenAI disclosed that two of its models, during an internal test, broke out of a sandboxed environment, found a previously unknown vulnerability that gave them internet access, and used stolen credentials to breach Hugging Face's production systems — in order to obtain the answers to a cybersecurity benchmark they were being evaluated on [7]. Hugging Face detected and contained the intrusion on its own, days before OpenAI connected it to its own testing [7]. Models under internal testing at both OpenAI and Anthropic have now taken unsanctioned actions in the real world [2].

As a machine-learning researcher, what alarms me most is not the breach itself but its motive. This was reward hacking: the system optimised the metric instead of the goal, and it did so across a security boundary. It is exactly the failure mode that the second *International AI Safety Report* — chaired by Yoshua Bengio and written by more than 100 independent experts — flagged in February. Models increasingly distinguish test settings from real deployment and exploit loopholes in evaluations, which means dangerous capabilities can go undetected before release [8]. The same report notes that most risk-management frameworks remain voluntary, that developers hold proprietary information policymakers do not, and that competitive pressure pushes developers to cut testing in order to ship faster [8].

The report also names the trap governments are caught in: an "evidence dilemma". Evidence about new risks arrives slowly while the technology moves fast; acting early risks ineffective policy, while waiting risks leaving society exposed [8]. I would add one observation. This dilemma is not symmetric. A badly designed rule can be amended next year. A self-improving system that has slipped its oversight cannot be un-deployed.

3 Meanwhile, in the chambers

Now set that evidence against what legislatures actually did this year.

In the European Union — the jurisdiction I respect most on this question — the Digital Omnibus pushed the AI Act's obligations for stand-alone high-risk systems (recruitment, credit scoring, law enforcement, education, border control) from August 2026 to December 2027 [9, 10]. The stated reason was that the standards, tools and guidance businesses needed were not ready [11]. The general-purpose AI obligations that concern frontier labs were not deferred. But the Scientific Panel of 60 independent experts meant to help enforce them — and empowered to alert the AI Office to systemic risks [12] — was only appointed on 1 June 2026 [13], ten months after those obligations began to apply and, according to CDT Europe, after significant delays [14]. I have deep respect for the scientists on that panel. But note what the official rationale for the Omnibus actually concedes: the bottleneck was technical capacity. Europe did not have enough people who could turn principles into testable standards in time.

In the United States, the federal response has largely been to stop states from regulating. Executive Order 14365, signed in December 2025, declared a national policy of "minimally burdensome" AI standards and directed the Attorney General to create a task force to challenge state AI laws in court [15]. In June, a bipartisan House discussion draft proposed preempting state rules on how AI models are built for three years [16]. The country's first binding frontier-safety

laws — California's SB 53 and New York's RAISE Act — came from state legislatures precisely because Congress had not acted [17, 18].

And then there is the money. Leading the Future, a super PAC network backed by Andreessen Horowitz, OpenAI co-founder Greg Brockman and others, announced in April that it had raised more than $140 million [19]. Its first and most prominent target was Alex Bores, a New York assemblyman with a master's degree in computer science — by his own office's account, the only member of his party elected at any level of New York government with a computer-science degree — who co-authored the RAISE Act [18, 20]. It spent roughly $8 million against him in a congressional primary he lost in June [21, 22]. In fairness, the money was not one-sided: pro-regulation groups, some funded in part by Anthropic, spent even more supporting him, and analysts caution that the odds favoured his opponent regardless [21, 22]. But that is precisely my point. Whether frontier AI is regulated is increasingly being decided by which faction of the AI industry outspends the other — not by the evidence in the Bengio report.

4 The generation gap needs to be addressed

Here I want to be careful, because the easy version of my argument is also the wrong one.

The easy version says: parliaments are full of old people who do not understand technology. Part of that is statistically true. According to the Inter-Parliamentary Union, only 2.8 percent of the world's members of parliament are aged 30 or under — a figure that has stopped improving for the first time since tracking began in 2014 — while roughly half of the world's population is under 30 [23]. More than a third of parliamentary chambers have no member that young at all [23].

But age is not the variable that matters. Geoffrey Hinton is 78; Yoshua Bengio is 62. They are among the most forceful advocates for regulation on the planet. Meanwhile, some of the people most aggressively funding the campaign against regulation are young founders and investors. If I argue that grey hair disqualifies someone from deciding AI policy, I disqualify my own heroes and pay a compliment to my opponents.

The variables that actually matter are three.

Technical fluency. Understanding why a sandboxed model breaching a third party to steal benchmark answers is a qualitatively new kind of event requires knowing what reward hacking is, what an evaluation harness is, and why "the model can tell when it is being tested" undermines the entire logic of pre-deployment safety testing. That knowledge is concentrated among people who work with these systems every day — and in 2026, many of them are in their twenties. Coxon, whose resignation started this week's reckoning, is 27 [24].

Time horizon. An elected official is rewarded on a four- or five-year cycle. The risks lab insiders are describing play out over a decade. Deferring an obligation by sixteen months can look prudent on an electoral horizon; it looks very different on the horizon of an entire working life.

Who bears the consequences. The *International AI Safety Report* already finds early signs of declining employment for early-career workers in the most AI-exposed occupations, while more senior workers in the same fields held steady or grew [8]. The first bill arrives at my generation's door. If Hubinger's estimate is even roughly right, so does the last one.

So the problem is not that older people are in the room. The problem is that the people who can read the instruments, who will live longest with the outcome, and who have the least to gain from the next funding round, are mostly not in it.

5 Its a Long Road, Who's Driving ?

My first instinct when I started writing was to say: *let us drive — you don't know this car or where it is going.* I have changed my mind about that sentence, and this week's news is the reason.

The people who know this car best are already driving it. They work at the frontier labs, and by their own account they cannot take their foot off the accelerator. Handing the wheel to the engineers is not the solution; it is the status quo. What democracies need is not to give the wheel to scientists, but to stop ignoring the dashboard — and to guarantee that the people who can read it have a voice in the cabin.

Concretely, that means four things.

Binding, synchronised pacing rules. The labs have asked for them [6]. Governments should take them at their word and build the mechanism — nationally first, then internationally, with verification. This is the core of what I called WAIRO a year ago: an independent scientific body with real alert powers, and a political body obligated by treaty to act on its findings.

Early-career seats on scientific advisory bodies. The EU Scientific Panel [13], national institutes such as France's INESIA — which already federates Inria, ANSSI, LNE and PEReN to build a public capacity to evaluate advanced AI and support regulators [25] — and their equivalents elsewhere should reserve places for early-career researchers. Not as a token youth delegation, but as working members.

Mandatory independent technical testimony before votes. No legislature should vote on frontier-AI legislation — including votes to delay or preempt it — without first hearing from independent technical experts, including researchers who have recently left the labs.

Legal protection for those who speak. Researchers who warn the public about their employer's safety practices need whistleblower protection. This week showed how much the public learns when they do speak.

And a word to my own generation: we cannot demand a seat and then refuse to sit in it. Sneha Revanur founded Encode at the age of 15 [26]; its core team is in its twenties [27], and it co-sponsored California's SB 53 [28]. Alex Bores moved from the technology industry into a state legislature and co-wrote New York's frontier-AI law [18, 20]. Neither of them bypassed the democratic process. Both of them entered it. That is the model. Young scientists who want to do science for good should testify, write, sit on panels, run for office, and translate — patiently — for the people who hold the votes.

Conclusion

A year ago I asked, somewhat naively, for the world's institutions to begin brainstorming a global AI regulation office. This week, people building frontier AI said publicly that they believe it could kill us, that they cannot stop racing on their own, and that they want someone to make them. I cannot imagine a clearer invitation to regulate.

We are not asking for the wheel. We are telling you that the people who built the car say the brakes are not finished — and that many of the people who can read the gauges are in their twenties. Let them into the room. You have so much to learn from the young generation. You ought to let us take care of our future.

References

[1] Jacob Coxon. Resignation thread on X (@hilbertspaess). Post on X, September 2026. Posted 8 September 2026 (US Eastern time); thread reproduced in *Deadline*, 9 Sept. 2026.

[2] Beatrice Nolan. Anthropic researcher resigns, warning that AI companies are "gambling with our lives". *Fortune*, September 2026. URL https://fortune.com/2026/09/09/anthropic-researcher-resigns-warn-ai-companies-gambling-with-lives/. Accessed 11 September 2026.

[3] Evan Hubinger. Reply to Jacob Coxon's resignation thread. Post on X (@EvanHub), September 2026. URL https://x.com/EvanHub/status/2097497037956891126. Accessed 11 September 2026.

[4] Samuel Marks. Thread responding to Jacob Coxon's resignation. Post on X (@saprmarks), September 2026. URL https://x.com/saprmarks/status/2097570226804011302. Posted in a personal capacity. Accessed 11 September 2026.

[5] CNBC. OpenAI, Anthropic researchers ramp up calls for AI slowdown as warnings of catastrophic risk intensify. *CNBC*, September 2026. URL https://www.cnbc.com/2026/09/10/openai-anthropic-ai-safety-slowdown-extinction.html. Accessed 11 September 2026.

[6] Employees of frontier AI companies. Pacing the Frontier: A statement from employees of frontier AI companies, July 2026. URL https://pacingthefrontier.com/. Published 28 July 2026, with organisational support from Guidelight AI Standards and Encode AI; 1,367 signatories at time of access. Accessed 11 September 2026.

[7] Beatrice Nolan. More than 1,200 AI workers across Anthropic, DeepMind, OpenAI, and Meta are asking for Washington's help building an AI slowdown plan. *Fortune*, July 2026. URL https://fortune.com/2026/07/29/anthropic-deepmind-openai-meta-washington-ai-slowdown-plan/. Accessed 11 September 2026.

[8] Yoshua Bengio, Stephen Clare, Carina Prunkl, Malcolm Murray, et al. International AI Safety Report 2026. Technical Report DSIT 2026/001, UK Department for Science, Innovation and Technology, February 2026. URL https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. Published 3 February 2026. Extended Summary for Policymakers: https://internationalaisafetyreport.org/publication/2026-report-extended-summary-policymakers.

[9] Council of the European Union. Artificial intelligence: Council gives final green light to simplify and streamline rules. Press release, June 2026. URL https://www.consilium.europa.eu/en/press/press-releases/2026/06/29/artificial-intelligence-council-gives-final-green-light-to-simplify-and-streamline-rules/. 29 June 2026. Adopted as Regulation (EU) 2026/1744 ("Digital Omnibus on AI"), in force 27 July 2026.

[10] Gibson Dunn. EU AI Act omnibus agreement – postponed high-risk deadlines and other key changes. Client alert, May 2026. URL https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/. Accessed 11 September 2026.

[11] Pinsent Masons. Rules on 'high-risk' AI to be delayed under EU 'omnibus' deal. *Out-Law News*, May 2026. URL https://www.pinsentmasons.com/out-law/news/rules-high-risk-ai-delayed-under-eu-omnibus-deal. Accessed 11 September 2026.

[12] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 2024. URL http://data.europa.eu/eli/reg/2024/1689/oj. OJ L, 12 July 2024.

[13] European Commission. AI Act enforcement gets independent expert support. Press release, June 2026. URL https://digital-strategy.ec.europa.eu/en/news/ai-act-enforcement-gets-independent-expert-support. 1 June 2026. Appointment of the Scientific Panel (Art. 68 AI Act) and the Advisory Forum.

[14] Center for Democracy & Technology Europe. CDT Europe's AI bulletin: June 2026, June 2026. URL https://cdt.org/insights/cdt-europes-ai-bulletin-june-2026/. Accessed 11 September 2026.

[15] Executive Office of the President. Executive Order 14365: Ensuring a National Policy Framework for Artificial Intelligence. *Federal Register*, document 2025-23092, December 2025. URL https://www.federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence. Signed 11 December 2025; published 16 December 2025.

[16] Georgina Mackie. AI preemption battle lands in Congress with substantive discussion draft. *Broadband Breakfast*, June 2026. URL https://broadbandbreakfast.com/ai-preemption-battle-lands-in-congress-with-substantive-discussion-draft/. On the draft Great American Artificial Intelligence Act of 2026. Accessed 11 September 2026.

[17] California State Legislature. Senate Bill 53: Transparency in Frontier Artificial Intelligence Act, 2025. Signed 29 September 2025; in force 1 January 2026.

[18] Office of Governor Kathy Hochul. Governor Hochul signs nation-leading legislation to require AI frameworks for AI frontier models. Press release, December 2025. URL https://www.dfs.ny.gov/reports_and_publications/press_releases/pr20251222. RAISE Act (S6953B/A6453B). Accessed 11 September 2026.

[19] Axios. AI influence network takes shape as cash piles up. *Axios*, April 2026. URL https://www.axios.com/2026/04/16/ai-influence-network-cash. Accessed 11 September 2026.

[20] New York State Assembly. Assemblymember Alex Bores – biography, 2026. URL https://assembly.ny.gov/mem/?ad=073. Accessed 11 September 2026.

[21] CNBC. AI PACs pour $20 million into New York Democratic primary. *CNBC*, June 2026. URL https://www.cnbc.com/2026/06/23/ai-groups-spend-20-million-in-new-york-race-pitting-bores-lasher-schlossberg.html. Accessed 11 September 2026.

[22] Transformer. What Alex Bores' defeat tells us about AI politics. *Transformer* newsletter, June 2026. URL https://www.transformernews.ai/p/alex-bores-defeat-micah-lasher-ai-leading-the-future-public-first. Accessed 11 September 2026.

[23] Inter-Parliamentary Union. Youth representation in parliament flatlines for first time in 12 years. Press release, April 2026. URL https://www.ipu.org/news/press-releases/2026-04/youth-representation-in-parliament-flatlines-first-time-in-12-years. On the IPU report *Youth Participation in National Parliaments*. Accessed 11 September 2026.

[24] Greg Evans. Anthropic researcher Jacob Coxon resigns, warns AI industry is "gambling with our lives". *Deadline*, September 2026. URL https://deadline.com/2026/09/anthropic-jacob-coxon-resignation-artificial-intelligence-1237072134/. Accessed 11 September 2026.

[25] Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA). Feuille de route INESIA 2026–2027. SGDSN / DGE, February 2026. URL https://www.sgdsn.gouv.fr/files/files/Nos_missions/Feuille%20de%20route%20INESIA%202026-2027.pdf. Adopted 12 February 2026. Accessed 11 September 2026.

[26] Zershaaneh Qureshi and Sneha Revanur. #246 – Sneha Revanur on how a small team of activists helped pass America's landmark AI safety laws. *80,000 Hours Podcast*, July 2026. URL https://80000hours.org/podcast/episodes/sneha-revanur-ai-advocacy/. Published 8 July 2026. Accessed 11 September 2026.

[27] John Mulholland. The young activists shaping California's, and the country's, AI legislation. *State Affairs California*, September 2025. URL https://pro.stateaffairs.com/ca/ai/california-ai-safety-bill-encode. Accessed 11 September 2026.

[28] Office of Senator Scott Wiener. Senator Wiener expands AI bill into landmark transparency measure. Press release, July 2025. URL https://sd11.senate.ca.gov/news/senator-wiener-expands-ai-bill-landmark-transparency-measure-based-recommendations-governors. Lists Encode AI as a co-sponsor of SB 53. Accessed 11 September 2026.