How AI Companies Want to Be Governed, and What That Means for AI Governance

techpolicy.press/how-ai-companies-want-to-be-governed-and-what-that-means-for-ai-governance

Jason Pielemeier, Jhalak M. Kakkar, Elonnai Hickok, Sudisha Mishra, Revati Vaidya November 2, 2026

Jason Pielemeier, Jhalak M. Kakkar, Elonnai Hickok, Sudisha Mishra, Revati Vaidya / Nov 2, 2026



President Trump, left, speaks with Meta CEO Mark Zuckerberg as Dario Amodei, CEO and co-founder of Anthropic addresses reporters outside the White House, Sept. 29, 2026, in Washington DC. (AP Photo/Alex Brandon)

Leading frontier AI companies, their employees, AI researchers, and governments have all become increasingly vocal about the need to regulate “frontier” AI. But how do leading AI companies envision public governance? By examining the recent pronouncements of the five most vocal major AI model developers — Anthropic, Google (a subsidiary of Alphabet), Meta, Microsoft, and OpenAI — a close reader can better answer this question. .

Our research, as part of the Multistakeholder Approach to Participation in AI Governance (MAP-AI), summarized below, aims to deepen understanding of, and shape conversations around, the relationships, responsibilities, and recourse needed to build meaningful, appropriate, sustainable, and democratically aligned AI governance. This is an initial compilation and overview of relevant proposals and positions, which will be updated over time. The documentation and comparative analysis supporting this analysis is available here.

A variety of proposals without clear rules

Reading through this recent flurry of “urgent” calls for regulation and the corresponding pushback from some competitors, advocates, and governments, helps provide clarity on what leading companies are asking for and what is most urgently needed — namely, governance that is inclusive, grounded in human rights, and incentivizes responsible model development and deployment by clearly and appropriately assigning responsibility, liability, and accountability.

The chart below summarizes our findings, which are described in detail by the rest of this essay.

How Five AI Companies Want to Be Governed

Positions in each company's major AI policy statements, mapped to the three areas examined in this analysis.

Anthropic

A new agency [with] the explicit power to test frontier models [and] block a model's release if it poses a significant risk of catastrophic harm. Global coordination around a standards-setting body that can control pausing or slowing down model release.
The most detail on how to ensure independence among assessors[, including] random assignment. The most extensive approach to transparency. A 15-day window to report stolen weights. Mandated, industry-wide whistleblower protections.
Opposes federal "preemption" [of] state law unless a federal law is passed that meets or exceeds the standards set out in their proposal. Areas for possible coordination with China[, including] a SALT-like treaty for slowing recursive self improvement.

OpenAI

More cautious on this front. Proposes that the existing Center for AI Standards and Innovation (CAISI) be built up as the government's central testing body, but without addressing whether it should be empowered to halt model releases.
Independent audits [are] one of three things every state law should require, alongside a public safety framework and incident reporting. Broader approaches that incorporate additional risks. Mandated, industry-wide whistleblower protections.
"Reverse federalism" model treats state laws as a bottom-up bridge to an eventual national standard, reserving national-security and highly technical review for the federal government and leaving states to lead on child safety, energy, and education.

Google

A new industry-funded, government-supervised body called the Federal AI Regulation Organization (FARO). Developers would self-certify to FARO before a "significantly more capable" release[, framed as] certification rather than a discretionary government power to block.
Annual audits through established audit firms, who would initially focus on whether a company followed its own stated process[, with] three procedural protections for the companies being audited. Broader approaches that incorporate additional risks.
Appears to defer to federal authority on national security without taking an explicit preemption position. Joined a July 2026 statement [endorsing] open weight models.

Microsoft

A new regulatory body [to] license the model for deployment, and license data centers for training. “Safety brakes” that let critical-infrastructure operators disengage a misbehaving system.
Arguing for “certain obligations to know one’s cloud, one’s customers, and one’s content” (KY3C). Tie outside testing to the proposed licensing scheme rather than a freestanding audit regime. Annual transparency reports.
Brad Smith has criticized the current patchwork as “regulation without transparent or complete rules.” Acknowledges a genuine, unresolved tension between security and openness. Advance notification of large training runs.

Meta

Rejects the premise of pre-release review. Proposes that AI companies provide partly-trained models and technical staff to the government while training is still underway. Delaying a release “even by a month” could cost the US its lead.
Does not reference external audits at all. Quiet on transparency reporting. The only company that explicitly calls for cooperation with law enforcement [on security incidents].
A “balance of power that favors individuals” focused on accelerating the diffusion of superintelligence. Focuses exclusively on encouraging cooperation between frontier labs and the US government, without addressing if or how international cooperation should take place.

Pacing

The most consequential disagreement among the proposals reviewed is about whether to slow down or stop frontier development. The call for “pacing the frontier” was most clearly stated in September by Anthropic CEO Dario Amodei in his aptly titled essay, “We Must Pace the Frontier.” This post echoes and elaborates upon Anthropic’s earlier “Advanced AI Framework” and its “Policy on the AI Exponential,” which ask Congress to create a new agency modeled on the US Federal Aviation Administration, with the explicit power to test frontier models, oversee independent evaluations, and block a model’s release if it poses a significant risk of catastrophic harm even after mitigations. In the most recent proposal, Amodei also suggests global coordination around a standards-setting body that can control pausing or slowing down model release after a thorough risk assessment.

While OpenAI CEO Sam Altman was among others who endorsed Amodei’s September call for pacing the frontier, his company’s public positions have heretofore been more cautious on this front. Its “Public Policy Agenda” and Chris Lehane’s “reverse federalism” essay ask US states to pass matching laws that federal

legislation would later absorb into a harmonized national approach. OpenAI proposes that the existing Center for AI Standards and Innovation (CAISI) be built up as the government's central testing body, but without addressing whether it should be empowered to halt model releases. It also nods toward the need for global coordination, suggesting "a US-led international forum that establishes accepted standards, provides expert and impartial analysis of capabilities and risks, and makes the technology available to nations and companies that participate and follow the rules," something Altman had previously suggested could be similar to the International Atomic Energy Agency.

Alphabet's chief scientist and Google DeepMind co-founder Demis Hassabis has also endorsed Amodei's call for pacing the frontier. In its more developed governance proposals, Google splits the problem in two: presenting frontier AI as a national-security question, and addressing everyday AI from a consumer-protection perspective. For frontier models, its "Pragmatic Approach to AI Governance" proposes a new industry-funded, government-supervised body called the Federal AI Regulation Organization (FARO), modeled on the Financial Industry Regulatory Authority (FINRA) or the North American Electric Reliability Corporation (NERC); Hassabis has also independently articulated his own thoughts on what FINRA-style regulation could look like. Developers would self-certify to FARO before a "significantly more capable" release, giving the regulator the ultimate power to intervene where necessary, although Google is careful to frame the approach as certification rather than a discretionary government power to block model release.

Microsoft's 2023 Blueprint acknowledges the need for new laws to address highly capable AI foundation models, making the case for a new regulatory body, which would ensure models comply with existing laws, license the model for deployment, and license data centers for training and deployment of models based on security, export controls, and safety protocols. It also proposes "safety brakes" that let critical-infrastructure operators disengage a misbehaving system. To this end, Microsoft committed to sharing specialized knowledge about its foundation models and supporting governments in defining requirements for licensing and enforcement. Microsoft has articulated further positions on pacing and other AI safety commitments in its recently published draft "Humanist AI Code of Conduct."

Meanwhile, Meta rejects the premise of prerelease review. Mark Zuckerberg's "The Future is for Everyone" letter argues that no single company or government should be trusted to set a fixed review timeline for everyone and instead proposes that AI companies provide partly trained models and technical staff to the government while training is still underway, giving officials, in his words, "an army of capable engineers" to find problems in real time rather than checking a finished product. He frames this approach as more protective of open, decentralized power than a licensing regime,

and warns that delaying a release “even by a month” could cost the US its lead in the global AI race.

Development, deployment, and disclosure

The companies’ proposals also converge on the high-level outline of what should be done to identify and address deployment-related risks: publish a safety framework, test models against a set of risk domains, and make disclosures about the results. A similar approach has also been articulated by a range of countries through a statement led by the governments of Finland and Norway. Beyond this basic testing and disclosure approach, differences emerge regarding the appropriate level of independence for testing, the range of risks to address, and the extent to which transparency should be mandated.

Anthropic gets into the most detail on how to ensure independence among assessors, proposing the development of a professional, AI-specific evaluation industry, with published qualification standards, dedicated funding so evaluators aren’t paid by the companies they assess, government ratings of evaluator quality, and random assignment so companies can’t shop for an easy reviewer. It pairs this with a 15-day window to report stolen weights, realized harm from a named risk, or a model that deceives its own developers to evade oversight.

OpenAI and Google both support independent audits in principle, and OpenAI names them as one of three things every state law should require, alongside a public safety framework and incident reporting. Google’s FARO would run annual audits through established audit firms, which would initially focus on whether a company followed its own stated process, with an eye toward more substantive reviews occurring once standards mature, alongside three procedural protections for the companies being audited (standard document sets, a chance to fix problems before the report is final, and confidential delivery).

Microsoft bases its Blueprint on the know-your-customer-approach to financial risk management, adapting it to the AI context and arguing for “certain obligations to know one’s cloud, one’s customers, and one’s content” (KY3C). This approach would tie outside testing to the proposed licensing scheme rather than a freestanding audit regime. By contrast, Meta does not reference external audits at all.

There is also substantial ambiguity across these proposals regarding what these evaluation and audit frameworks would focus on. While every company proposal reviewed recognizes “existential” harms as a necessary focus for risk management, Google, OpenAI, and Microsoft offer broader approaches that incorporate additional

risks, including deepfakes, privacy concerns, as well as energy, environmental, and employment-related issues.

Transparency is a cornerstone of responsible business conduct, enabling third parties to hold companies and governments accountable. Over the past two decades, transparency reporting has evolved into a robust, established practice on digital platforms. Anthropic outlines the most extensive approach to transparency, noting that companies should publish safety frameworks, issue a risk report every six months, release a system or model card when more capable models are released, and provide certification of compliance with evaluation requirements. OpenAI, Google, and Microsoft also support publishing some form of overarching policy outlining key practices, though what it should contain differs from company to company. Microsoft proposes annual transparency reports and supporting a national registry of high-risk AI systems. In contrast, Zuckerberg's post is quiet on transparency reporting.

All five companies nod to the importance of reporting security incidents, though there is disagreement and ambiguity over what would constitute a reportable incident, how soon the incident should be reported, to whom it should be reported (the public, the regulator, the data center), as well as what should be reported. Notably, Meta is the only company that explicitly calls for cooperation with law enforcement in these cases.

Anthropic and OpenAI also both push for mandated, industry-wide whistleblower protections, while the other companies are silent on this, even though Meta's and Microsoft's internal safety frameworks (their Frontier AI Framework and Frontier Governance Framework, respectively) include anonymous reporting channels and anti-retaliation commitments for their own employees.

Collective action

Collective action challenges sit at the core of AI governance, and the tension between competition and coordination is evident across the companies' proposals, with one notable exception: all five companies appear eager to develop predictable, confidential processes through which they can disclose risk-related information to the US government.

As a starting point, all five US companies target their proposals to the US political and regulatory context, with minimal references to the regulatory developments in the EU and elsewhere. While this is not surprising given that all five are US-headquartered companies, it nevertheless shows that companies may not sufficiently appreciate the fact that other democratic countries are unlikely to defer completely to regulatory

approaches modeled on US institutions and enforced exclusively through the US legal system. These recent calls for regulation should also be interpreted in light of documented instances in which some of these same companies have actively sought to undermine regulation in the US.

Where the companies do address the international dynamics around AI governance, they articulate a shared desire for interoperability with allies, either implicitly or explicitly leaving some countries out, while juxtaposing the feasibility of these efforts with the need to stay ahead of China specifically.

Meta stands out for favoring a "balance of power that favors individuals" focused on accelerating the diffusion of superintelligence. After committing to "contribute to this by building leading personal superintelligence agents and models, including open source models, and delivering them to billions of people around the world," Zuckerberg states that "[a]ny policy that slows American model releases — even by a month — could add significant risk to American leadership while letting foreign models race ahead," and focuses exclusively on encouraging cooperation between frontier labs and the US government, without addressing if or how international cooperation should take place.

While Anthropic's earlier proposals focused on global coordination among democracies, Amodei's most recent proposal goes further and proposes areas for possible coordination with China specifically. While acknowledging the likely limits on such cooperation, he proposes several areas for possible agreement, including prohibiting certain dangerous uses of AI, mutual commitments to prerelease testing, a SALT-like treaty for slowing recursive self-improvement, and finally, and least realistically, mutual commitment to a full-scale pause.

Within the US context, the companies have staked out diverging positions on the question of "federalism" — i.e., whether activities should be regulated exclusively and uniformly through federal law, or through potentially divergent state laws. Anthropic opposes federal "preemption" (superseding) of state law unless a federal law is passed that meets or exceeds the standards set out in their proposal. Most recently, according to reporting by Politico, a draft bill circulated by the company would allow preemption "related to public disclosure of risk, third-party evaluation and reporting of critical incidents."

By contrast, OpenAI's "reverse federalism" model treats state laws as a bottom-up bridge to an eventual national standard, reserving national-security and highly technical review for the federal government and leaving states to lead on child safety, energy, and education. Google appears to defer to federal authority on national security without taking an explicit preemption position, while Microsoft's Brad Smith

has criticized the current patchwork as “regulation without transparent or complete rules,” and Meta’s documents are silent on the question entirely.

The companies also take somewhat different approaches to the issues raised by open-weight models. Not surprisingly, given its focus on open weights, Meta bases its proposed approach on openness as an antidote to concentrated power, arguing that restricting today’s open ecosystem “would be a mistake.” Microsoft acknowledges a genuine, unresolved tension between security and openness and joined a July 2026 statement from a wide range of companies, including Google and Meta, endorsing open weight models as an important part of a strong, open ecosystem that diffuses AI across the US economy. Meanwhile, Anthropic and OpenAI are missing from that list and do not address the question in their documents, beyond Anthropic’s stated concern with detecting and reporting attempts to copy or steal proprietary weights.

All five companies converge on the idea that the US government should get some form of privileged early information about and in some cases access to frontier capabilities, whether through incident reports routed to federal agencies and national labs (Anthropic), a federal testing framework that gets tools to critical-infrastructure defenders quickly (OpenAI), early notice for models that break new ground on sensitive national-security capabilities (Google), advance notification of large training runs, comprehensive risk assessments focused on dangerous or breakthrough capabilities, and extensive prerelease testing by internal and external experts as part of licensing requirements (Microsoft), or, most dramatically, handing over partially trained models and staff during training itself (Meta).

Finally, Amodei’s recent proposal calls for proactive safety-focused cooperation between leading US AI companies and calls for the government to mediate or enable those discussions “for antitrust reasons.” The idea of issuing such an antitrust waiver, which has also been supported by OpenAI and Google, has been introduced in the US Congress and has already provoked a backlash.

Observations and recommendations

These company proposals offer useful building blocks for developing regulatory institutions, evaluation mechanisms, safety frameworks, and information-sharing mechanisms. However, they are largely company-focused, US-centric, and anchored to technical or model-level safety. They leave open fundamental questions such as who should govern AI, in whose interests, for which context, and with what degree of independence; what mechanisms should exist for accountability and participation; and how domestic and international governance and cooperation should interrelate.

As the MAP-AI community has recently set forth, rather than working toward a comprehensive treaty and a single global regulator, it is critical to develop multiple, intersecting, and coherent mechanisms to effectively govern AI. This ecosystem approach should be polycentric and people-centered. Coherence should be built with the norms and principles established through efforts like the Council of Europe's Framework Convention on Artificial Intelligence, the UNESCO Recommendation on the Ethics of Artificial Intelligence, the OECD Principles for Trustworthy AI, the Freedom Online Coalition's Joint Statement on AI and Human Rights, and the UN Office of the High Commissioner for Human Rights' Foundational Paper on Advancing Responsible Development and Deployment of Generative AI.

The authors have written this analysis in their capacities working on the MAP-AI project; nothing here should be interpreted as representing institutional positions of CCG, GNI, or GNI members.



Support Tech Policy Press
If you've found our work helpful, consider supporting us.

Authors



[Jason Pielemeier]

Jason Pielemeier is the Executive Director of the Global Network Initiative (GNI), a dynamic multi-stakeholder human rights collaboration, building consensus for the advancement of freedom of expression and privacy amongst technology companies, academics, human rights and press freedom groups and so...

[Jhalak M. Kakkar]

Jhalak M. Kakkar is Executive Director at the Centre for Communication Governance at National Law University Delhi as well as a Visiting Professor at the National Law University Delhi. She leads the academic and policy research at CCG across pressing information law and policy issues such as data go...





[Elonnai Hickok]

Elonnai Hickok is the Managing Director at Global Network Initiative. Elonnai Hickok has contributed to the digital rights field for 15 years through international policy initiatives. Elonnai serves as a current co-chair of the Freedom Online Coalition Advisory Network, and has researched and writte...



Sudisha Mishra

Sudisha Mishra is a Programme Officer at the Centre for Communication Governance at National Law University Delhi (CCG). Her current focus and research interests lie at the intersection of global AI governance, and global south solidarity and leadership at multistakeholder AI forums. Prior to this, ...



Revati Vaidya

Revati Vaidya is a Master in Public Administration candidate at Harvard Kennedy School and a Research Assistant with the Ash Center’s Informed Public Project. As a 2026 Summer Fellow with the Global Network Initiative, she contributed to its Multistakeholder Approaches to AI (MAP-AI) project. Her in...

Topics

Related