Lab Statements on AI Governance

Purpose of this document: Overview of public statements on AI governance proposals issued by frontier AI labs (OpenAI, Google DeepMind, and Anthropic)
Status: 9 person-hour screening for public commitments on governance regimes and policies by Kevin Wei, Richard Moulange, Lisa Soder, and Lennart Heim
Shared with: this copy is shared publicly

A number of leading AI companies, including OpenAI, Google DeepMind, and Anthropic, have voiced their support for AI governance proposals. These proposals include, for instance:

- **Evaluations of advanced AI systems** throughout their lifecycle (e.g. pre-training, pre- & post-deployment evaluations by independent auditors)
- **Coordination on safety-relevant aspects** between AI labs and also between national governments
- **Development of shared / public standards** for AI safety and for model evaluation
- All three labs have also already implemented **system cards**, **usage policies**, and various **cybersecurity measures**

Legend: green cells indicate the AI lab supports the proposal, while blue cells indicate the AI lab could support the proposal (or supports a similar proposal). Blank cells indicate we did not find any public information indicating that the AI lab supported (or did not support) the proposal – it may be that a longer search would have produced relevant material. Where proposals may require new government legislation, it is indicated in the relevant cell.

Proposal	Proposal Details	Support Count[1]	Currently Deployed?[2]	OpenAI	DeepMind / Google	Anthropic	Category of Ask
Safety coordination (between labs and/or modulated by national governments)	Coordination between AI labs (and/or governments coordinating between AI labs) on AI safety (e.g. capability limits, research, evaluation, etc.)	3		Yes: mentions coordinating on compute usage; White House voluntary commitment	Yes: White House voluntary commitment; also NTIA comment could refer to coordination between countries or labs;	Yes: wants clarity on antitrust law; White House voluntary commitment	International governance
Coordination between governments on standards		3		Yes: IAEA	Yes: NTIA (p. 4, 31)	Yes	International governance
Multi-stakeholder processes to develop standards, requirements, etc.		3		Yes	Yes: also NTIA (p. 4)	Yes	International governance
Risk-based regulation	Regulation based on the level of risk posed by AI models	3		Yes: supports licensing / eval requirements	Yes: NTIA (p. 3, 7, 19)	Yes: risk thresholds, "risk responsive evaluations"	International governance

[1] Number of labs that support the policy
[2] Currently deployed by labs that support the policy

Proposal	Proposal Details	Support Count[1]	Currently Deployed?[2]	OpenAI	DeepMind / Google	Anthropic	Category of Ask
				specifically for models beyond a certain capability threshold			
International interoperability for AI standards	Ensure safety standards and regulations are consistent and aligned between jurisdictions	3		Yes	Yes	Yes	International governance
Improved cybersecurity	Various different proposals to improve cybersecurity for AI models and/or AI products	3	Yes (but more can be done)	Yes: currently providing funding for this; White House voluntary commitment	Yes: NTIA p16 (access controls); White House voluntary commitment	Yes; White House voluntary commitment	Lab requirement
Post-deployment evals/standards		3		Yes; also in Frontier Model Forum objectives	Yes; also in Frontier Model Forum objectives	Yes: p 13, risk thresholds also in Frontier Model Forum objectives	Lab requirement
Usage Policies	Labs disallowing their models/products being used in specific use cases in their Terms of Use / Terms of Service	3	Yes	Yes	Yes	Yes: supports mandating that content as AI generated where applicable	Lab requirement
Pre-deployment independent audit/evals	Mandate that AI models be evaluated by an external/independent auditor prior to being made publicly available	3	Yes	Yes	Yes: also in NTIA and facial recognition case study here	Yes: "[eventually] mandatory evaluations"	Lab requirements
[Develop and fund] Safety and evaluation standards / best practices		3		Yes	Yes	Yes	Lab requirements
System Cards	Lab-created documentation containing information about the AI model, model use case, training data and process, evaluation data and metrics, and social implications / risks, in an accessible format for non-experts	3	Yes	Yes; also White House voluntary commitment	Yes: NTIA (p. 24); also White House voluntary commitment	Yes; also White House voluntary commitment	Lab requirement
Watermarking for audiovisual content	Develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content. This does not include content "readily recognizable as generated by a company's AI"	3		Yes: White House voluntary commitment	Yes: White House voluntary commitment	Yes: White House voluntary commitment	
Pre-deployment red-teaming		3	Yes	Yes; White House voluntary commitment	Yes: White House voluntary commitment	Yes: says these should be mandated (may need new law); White House voluntary commitment	Lab requirement
Bug bounty	Monetary payments, open to the public, to independent / external actors who discover security vulnerabilities (note: this proposal is limited to security vulnerabilities and not undesirable model outputs)	3	Yes	Yes; also White House voluntary commitment	Yes: White House voluntary commitment	Yes: White House voluntary commitment	Lab requirement

Proposal	Proposal Details	Support Count[1]	Currently Deployed?[2]	OpenAI	DeepMind / Google	Anthropic	Category of Ask
Pre-training independent review	Mandate that proposed AI models are checked by an external/independent auditor before models are trained	3		Yes	Supports internal reviews before training	Supports pre-training registration for large models	Lab requirement
[Governments should fund] Technical alignment solutions		2		Yes		Yes: specifically interpretability, bias	Government funding
[U.S.G. should] fully fund NIST	U.S. Congress should give NIST the budget it requested for the next FY	2			Supportive: NTIA p4 (emphasises NIST as hub of 'hub and spoke' regulatory model)	Yes	Government funding
Fund academic research / give academics more compute		2			Yes: Shevlane paper suggests investing in evaluation ecosystem, including for academic researchers (but doesn't mention compute specifically)	Yes	Government funding
Licensing and/or registration requirements for AI developers		2		Yes: IAEA		Yes: registry + pre-registration prior to large training runs (may need new law)	Lab requirement
Mandatory disclosure of evaluation / audit results		2		Yes		Yes: would need new law	Lab requirement
Training and education programs for AI-displaced workers		1		Yes			Government funding
[USG should] fund CHIPS Act		1				Yes	Government funding
AI testbeds		1				Yes: specifically run by NTIA	Government infrastructure
Centralized datasets for AI evaluation		1				Yes: specifically by NTIA	Government infrastructure
"IAEA" for AI / global central AI regulator		1		Yes: would need new law			International governance
Compute limit for new models		1		Yes			Lab requirement
Tracking AI inputs (compute, energy)		1		Yes			Lab requirement

Resources

- AI labs' statements on governance
- OpenAI
    - Governance of Superintelligence (blog post)
    - Planning for AGI and beyond (blog post)
    - Sam Altman U.S. Senate testimony
    - Comment on NTIA “AI Accountability Policy Request for Comment”
- Google
    - Google AI Principles
    - A Shared Agenda for Responsible AI Progress (blog post)
- DeepMind
    - Comment on NTIA “AI Accountability Policy Request for Comment”
    - An early warning system for novel AI risks (blog post and paper)
    - DeepMind Operating Principles (consistent with Google AI Principles)
- Anthropic
    - Charting a Path to AI Accountability (blog post)
    - Comment on NTIA “AI Accountability Policy Request for Comment”
    - “An AI Policy Tool for Today: Ambitiously Invest in NIST” (blog post and associated policy memo)
    - Comment on NIST "Study To Advance a More Productive Tech Economy"
- All of the above, plus Meta, Amazon, Inflection AI and Microsoft committed to the White House Voluntary Commitments (July 2023)
    - This was followed up by the Frontier Model Forum from OpenAI, Google, Microsoft and Anthropic