CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Summary of “AI Governance at the Frontier: Unpacking Foundational Assumptions”

Policymakers and researchers must contend with a variety of AI governance proposals amid great uncertainty about the future of AI development. **This report presents an analytic approach to derive AI governance proposals’ underlying assumptions, which are the foundational elements of a proposal that facilitate its success.** By adopting this approach, policymakers can take informed steps to govern AI systems while preserving future decision-making flexibility, and researchers can clarify points of divergence and convergence across AI governance proposals.

Our approach involves deriving unique and shared assumptions (see the table below for an example) across proposals by answering three questions:

1. What risks are important to mitigate and who should have primary oversight of frontier AI?
2. Who is delegated tasks and able to play a role?
3. Would the proposed mechanisms or tools actually achieve the proposal’s objectives?

We apply these questions to five U.S.-centric AI governance proposals from industry, academia, civil society, and the federal and state governments that are generally aimed at governing frontier AI systems. We find that **most proposals view AI-enabling talent and AI processes and frameworks as important enablers of AI governance.** However, proposals lack consensus regarding the techniques that are most effective at mitigating AI risks and harms.

Table 9. Shared Assumptions About Which Techniques or Categories of Techniques Are Effective

Which techniques or categories of techniques are effective?	OpenAI Proposal	Zero Trust AI Governance	Managing Emerging Risks to Public Safety	SB-1047	Framework to Mitigate AI-Enabled Extreme Risks
Preventing model leakage or theft	🟩			🟩	🟩
Watermarking or implementing content provenance techniques	🟩	🟩			
Attribution of harms from AI		🟩		🟩	
Monitoring compute access or identifying when users are training frontier AI				🟩	🟩
Identifying and tracking frontier AI risks before they become harms	🟩		🟩	🟩	🟩

Source: CSET.

Policy Considerations:

Our case study bears lessons that are broadly applicable to policymakers and other stakeholders seeking to analyze any proposal.

1. **Policymakers should leverage proposals' assumptions to more precisely understand disagreements and shared views among stakeholders.** Some proposals that initially appear incompatible may actually share assumptions that reflect common AI governance priorities. Conversely, similar-looking proposals may actually carry different assumptions regarding the appropriate mechanisms and actors to implement AI governance.

2. **Policymakers can take action in an uncertain and rapidly changing environment by addressing common assumptions across proposals.** Shared assumptions across AI governance proposals may represent practicable policy actions (such as establishing AI frameworks and a strong talent base) that accommodate a variety of different stakeholders and AI forecasts.

By adopting our analytic approach, U.S. policymakers and researchers alike can move away from rhetorical debates about AI governance and better prepare the United States for a range of possible AI futures.

For more information:

- Download the report: https://cset.georgetown.edu/publication/ai-governance-at-the-frontier/
- Contact us: cset@georgetown.edu