Issue Brief

AI Governance at the Frontier

Unpacking Foundational Assumptions

Authors
Mina Narayanan
Jessica Ji
Vikram Venkatram
Ngor Luong

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

November 2025

Executive Summary

As artificial intelligence diffuses throughout society, policymakers are faced with the challenge of how best to govern the technology amid uncertainty over the future of AI development. To meet this challenge, many stakeholders have put forth various proposals aimed at shaping AI governance approaches. **This report outlines an analytic approach to help policymakers make sense of such proposals and take steps to govern AI systems while preserving future decision-making flexibility.** Our approach involves analyzing common assumptions across various proposals (as these assumptions are foundational elements for the success of multiple proposals), as well as unique assumptions across individual 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 this analytic approach to five U.S.-centric AI governance proposals that originate from industry, academia, civil society, and the federal and state governments. These proposals are generally aimed at governing frontier AI systems, which possess cutting-edge capabilities and therefore pose some of the most challenging questions for AI governance. Our analysis reveals 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.

Our analysis also bears lessons that are broadly applicable to policymakers seeking to analyze any proposal. Our case studies demonstrate that **1) policymakers should leverage proposals' assumptions to more precisely understand disagreements and shared views among stakeholders** and **2) policymakers can take action in an uncertain and rapidly changing environment by addressing common assumptions across proposals**. By adopting our analytic approach, U.S. policymakers can move away from rhetorical debates about AI governance and better prepare the United States for a range of possible AI futures.

Table of Contents

Executive Summary 1
Introduction 3
Report Scope 5
Guiding Questions 5
Selection Criteria 6
Limitations 6
Proposal Case Studies 7
Industry: OpenAI’s Approach to Frontier Risk 7
Civil Society: AI Now Institute’s Zero Trust AI Governance 9
Academia: Frontier AI Regulation: Managing Emerging Risks to Public Safety 11
State Government: California’s Safe and Secure Innovation for Frontier Artificial Intelligence Models Act 13
Federal Government: Framework to Mitigate AI-Enabled Extreme Risks 17
Analysis 19
Deriving Assumptions 19
Shared Assumptions 20
What Risks Are Important to Mitigate And Who Should Have Primary Oversight of Frontier AI? 21
Who Is Delegated Tasks and Able to Play a Role? 22
Would the Proposed Mechanisms or Tools Actually Achieve the Proposal’s Objectives? 24
Unique Assumptions 30
Discussion 32
Conclusion 34
Authors 35
Acknowledgments 35
Appendix A: Assumption Interdependencies 36
Endnotes 37

Introduction

The rapid pace of artificial intelligence development has posed a significant challenge for AI governance. New AI capabilities are being publicly released faster than policymakers can react to them, creating an uncertain environment in which they must make decisions about how to govern AI. In addition to addressing direct harms from AI systems, U.S. policymakers at all levels of government must contend with potential national security risks, international competition, guardrails that advance AI safety without stymieing innovation, and allocation of resources for AI development.

Naturally, as AI becomes an increasingly pressing issue for policymakers, they are receiving input on how to govern it from a variety of stakeholders. These stakeholders have varied interests and motivations, and many play a significant role in AI governance. Policymakers must remain attuned to varied stakeholder inputs, but balancing these interests is no easy task due to the scope of problems that AI governance must grapple with. For instance, publicly available AI tools have worsened existing problems such as fraud, online harassment, algorithmic discrimination, and the proliferation of harmful content.[1] Simultaneously, AI capabilities present an array of relatively novel risks, such as individuals developing unhealthy emotional attachments to AI chatbots or AI models hallucinating inaccurate information.[2] Policymakers are attempting to navigate this complicated landscape while also trying to account for AI's potential benefits and the importance of remaining an international leader in AI.

Faced with a rapidly evolving technology and the need to balance many (sometimes competing) interests in AI, policymakers may be unsure about how to respond to the proliferation of AI governance proposals. We define an AI governance proposal as any proposal that seeks to harness the benefits and mitigate the risks of AI by governing the development (including inputs such as data or computing power) or deployment (including the use and management) of AI systems. Proposals can vary in detail and scope and may include recommendations such as filtering AI systems' outputs or constraining AI development through voluntary or mandatory mechanisms.

We believe a logical first step that policymakers can take when analyzing proposals is to identify their assumptions, i.e., the conditions that each proposal presumes to be true. By pinpointing assumptions, policymakers can gauge the feasibility of different proposals, identify areas that merit further investigation, and consider different opinions about the future of AI. In particular, assumptions that are shared across proposals may indicate opportunities for consensus-building among different stakeholders. These common assumptions are effectively enablers of the success of multiple proposals. If they hold true, common assumptions can enable an array of policy options; if untrue, they could undermine entire governance frameworks. By analyzing and, when appropriate, investing in these assumptions, policymakers can take action while also preserving decision-making flexibility down the road—an important consideration when operating under great uncertainty. Policymakers can also identify unique assumptions within proposals to familiarize themselves with new approaches or different perspectives, or note areas where consensus-building may be more challenging.

By analyzing and, when appropriate, investing in these assumptions, policymakers can take action while also preserving decision-making flexibility down the road—an important consideration when operating under great uncertainty.

This report presents an approach for analyzing common and unique assumptions from policy proposals. We analyze five AI governance proposals from industry, civil society, academia, state government, and the federal government—all of which have a U.S. focus and have shaped public discourse around AI governance. This paper was drafted before the proposal and passage of California State Bill (SB) 53, California's recently passed frontier AI law. As such, one of our proposals is SB-1047, a proposed bill that was related to SB-53 and more comprehensive in scope. These proposals are generally aimed at governing frontier AI systems, which possess cutting-edge capabilities and therefore pose some of the most challenging questions for AI governance. This report does not judge the merit of these proposals or advocate for a particular governance approach. Instead, it demonstrates the significance and usefulness of identifying the assumptions that underpin different AI governance proposals. We break down proposals into their key components using the four guiding questions enumerated in the Report Scope section: **why** govern, **what** to govern, **who** governs, and **how** to govern. The following Proposal Case Studies section applies the guiding questions to each proposal.

In the Analysis section, we use the guiding questions to identify shared and unique assumptions among the five proposals. Our analysis reveals 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. We also identify two key takeaways: 1) Policymakers should leverage proposals' assumptions to more precisely understand disagreements and shared views among stakeholders and 2) policymakers can take action in an uncertain and rapidly changing environment by addressing common assumptions across proposals.

Report Scope

Guiding Questions

The guiding questions in Table 1 break the AI governance proposals we examine into their key components, which support analysis of a proposal's assumptions. At minimum, a proposal should articulate a goal (why govern) and a way to accomplish that goal (what to govern, who governs, how to govern). These are key components of a proposal that should withstand scrutiny. Without the means to accomplish a goal, a proposal is merely an aspiration, and without a stated goal, a proposal is a prescription with no clear purpose. The question of when to govern was excluded from our analysis because proposals rarely discussed the sequencing of different governance mechanisms, although the timing of interventions is an important factor that can drive different AI governance approaches.

These key components should not only stand up to scrutiny on their own but should also work together. For example, the actors responsible for governing (who governs) should be capable of harnessing tools and mechanisms to govern (how to govern). This interaction of key components, among others, can be used to identify assumptions—a process that is described in greater detail in the Deriving Assumptions section.

Table 1: Guiding Questions

Question	Description	Examples
Why govern?	Objectives or reasons for governing	- Promote AI innovation - Protect humanity from catastrophic risks
What to govern?	Broad areas to be governed	- Test and evaluation procedures - AI competition
Who governs?	Actors that leverage tools or mechanisms to govern	- Congress - AI developers
How to govern?	Tools or mechanisms to govern	- Enforce existing laws

		● Map risk profiles to frontier AI models

Source: CSET.

Selection Criteria

We apply the guiding questions in Table 1 to identify the key components of five AI governance proposals in the Proposal Case Studies section. These proposals were selected based on the following characteristics:

- **Govern frontier AI models:** The proposals in this report are generally aimed at governing frontier AI systems.* Frontier AI systems possess capabilities that are "comparable to or slightly beyond the current cutting-edge."[3] Given that these systems represent the current state of the art, they pose some of the most important and challenging questions about the benefits and risks of AI. Therefore, establishing effective governance of these systems could be particularly impactful.

- **Prescribe concrete policy actions:** The proposals in this report offer concrete policy recommendations, suggesting clear actions that policymakers, AI developers, and other stakeholders can take to govern AI systems. We can more readily identify key components from concrete recommendations than from general guidance that does not specify particular actors or actions.

- **Sourced from different stakeholders:** The proposals in this report are sourced from different stakeholder groups. Although the groups may share views and their members may overlap, they collectively represent a spectrum of views on AI governance from industry, civil society, academia, and government.

Limitations

Our case studies are not intended to be representative of their respective sectors or industries. There is a fair amount of disagreement within each stakeholder group (e.g., between individual companies, researchers, and policymakers), and while our analysis will highlight some similarities and differences between specific proposals, these will necessarily omit viewpoints that are not reflected in these case studies. For example, the proposals assessed in this report lack many international perspectives on AI

* The one exception is the AI Now Institute's Zero Trust AI Governance proposal. See page 9 for a more detailed discussion of why this proposal was selected.

governance because the authors of the proposals are largely based in the United States. Furthermore, our case studies will not explore the precise details of every single proposal because our guiding questions are lightweight and widely applicable to a variety of proposals.

Proposal Case Studies

The case studies below summarize “OpenAI’s Approach to Frontier Risk,” the AI Now Institute’s “Zero Trust AI Governance” proposal, the preprint “Frontier AI Regulation: Managing Emerging Risks to Public Safety,” the proposed California bill Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (better known as State Bill 1047 or SB-1047), and the congressional “Framework to Mitigate AI-Enabled Extreme Risks.” These are the industry, civil society, academic, state government, and federal government proposals, respectively. We also apply guiding questions to each proposal to surface key components that provide the building blocks for our analysis of assumptions.

Industry: OpenAI’s Approach to Frontier Risk

In October of 2023, OpenAI published the company’s approach to frontier AI in response to the United Kingdom’s request for voluntary commitments to promote safety, security, and trust in AI at the AI Safety Summit.[4] The UK government proposed five principles that OpenAI endorsed, along with Google DeepMind, Microsoft, and Anthropic.[5] While OpenAI’s proposal for governing frontier AI addresses some of the principles and covered topics related to frontier risks that were raised at the AI Safety Summit, the company’s proposal largely reflects its internal approach to AI governance.

This proposal was selected because OpenAI is a key player in the AI industry and has built many of the most well-known and technically capable models. While some of the provisions in OpenAI’s proposal are no longer applicable, it still captures much of a leading AI developer’s thinking about internal AI governance.

Table 2: Key Components of OpenAI's Approach to Frontier Risk

Why govern: OpenAI needs to protect humanity from frontier AI risks, including catastrophic risks		
What to govern:	**Who governs:**	**How to govern:**
Frontier model development	OpenAI team and a joint Deployment Safety Board (DSB) with Microsoft	● Establish a governance structure (The Preparedness Framework) for evaluation, monitoring, accountability and oversight and decide on deployment rather than on earlier training decisions.
Research into frontier AI risks	OpenAI's Superalignment team and Preparedness team	● Scale efforts to align superintelligence and identify, track, and prepare for frontier risks.
Model test and evaluation procedures	External red-teamers, OpenAI's Red Teaming Network, Alignment Research Center* and developers	● Test the model (including through red-teaming), assess risks from power-seeking behavior, and publish system cards to increase transparency.
Model safety and security post-deployment	OpenAI team	● Detect abuse or unforeseen risks via security controls and monitoring and avoid external distribution of model weights.
Data inputs	OpenAI team	● Filter and remove unsafe training data.
Model outputs	OpenAI team	● Implement watermarking, classifiers, and metadata-based approaches.

Source: CSET.

* Since renamed to METR (Model Evaluation & Threat Research).

The primary objective of OpenAI's proposal is to manage frontier AI risks, and in particular, catastrophic risks. This 2023 proposal covers six main governance areas, including OpenAI's Preparedness Framework, which details their approach to tracking and mitigating catastrophic risks. The proposal also allocates responsibilities for governance to various OpenAI subteams and a select few partner organizations, including the Superalignment team responsible for aligning AI systems with human intentions (which has since disbanded) and an open network of red teamers from fields such as biology, political science, and sociology.[6] Unlike the other proposals, which are broader in scope, this proposal focuses primarily on OpenAI's internal governance policies.

Civil Society: AI Now Institute's Zero Trust AI Governance

The civil society organizations Accountable Tech, the AI Now Institute, and the Electronic Privacy Information Center jointly released an AI governance framework titled "Zero Trust AI Governance" in August of 2023.[7] The term "zero trust" in the title originates from cybersecurity, where it refers to the practice of constantly authenticating and verifying all users in a network instead of assuming that certain users can be trusted by default.[8] In the context of AI governance, the title suggests that the U.S. government has an obligation to constantly monitor and enforce rules governing the behavior of AI companies and developers.

This proposal was selected because it was co-authored by representatives from three civil society organizations that have been prominent advocates in the AI policy space. "Civil society" is a broad category encompassing a variety of stakeholders outside of the government and private sector, many of which advocate for different interests and hold opposing views. The Zero Trust AI Governance proposal offers a different perspective than the other proposals in its focus on consumer safety and competition policy. It also does not focus strictly on *frontier* AI systems, although many of its recommendations address use cases associated with generative or general-purpose AI systems.

Table 3: Key Components of AI Now Institute's Zero Trust AI Governance

Why govern: The federal government should mitigate harms from AI development and deployment by changing the incentive structure associated with AI developments		
What to govern:	**Who governs:**	**How to govern:**
Enforcement of existing laws	Federal government agencies, administrative bodies, and law enforcement	● Enforce existing anti-discrimination, consumer protection, and competition laws. ● Clarify the limits of Section 230 and support plaintiffs seeking redress of harms related to AI systems. ● Establish provenance, authenticity, and disclosure standards for generative AI systems.
AI use cases and training data	Congress	● Establish rules around the use of AI when necessary, including prohibiting certain unacceptable practices. ● Prohibit most secondary uses and third-party disclosure of personal data.
Competition in the AI industry	Federal Trade Commission	● Prevent toxic competition via structural interventions.
AI developers' behavior	AI developers and companies	● Affirmatively demonstrate compliance with pre- and post-deployment requirements throughout all stages of AI development. ● Uphold core AI ethics principles. ● Proactively notify users when AI is being used and maintain an easy complaint mechanism for users. ● Swiftly report any serious risks that have been identified. ● Grant third-party auditors full API and data access.
AI systems post-deployment	Third-party independent auditors with full API and data access	● Perform ongoing risk monitoring of AI systems post-deployment, including conducting independent audits that are publicized.

Source: CSET.

This civil society proposal is in some ways a response to policy proposals put forth by AI companies and industry advocacy groups. It emphasizes taking swift action, enforcing regulatory regimes, and shifting the burden of accountability from consumers—or those who suffer harm from AI systems—to AI companies and deployers. The proposal advocates for greater transparency from AI companies regarding their development practices and identification of risks, and encourages companies to grant third parties greater levels of access to their AI systems.

Academia: Frontier AI Regulation: Managing Emerging Risks to Public Safety

A group of 24 scholars submitted a version of the preprint "Frontier AI Regulation: Managing Emerging Risks to Public Safety" in July of 2023 and last revised the paper in November of 2023.[9] The paper was authored by individuals from think tanks and other nonprofit research institutions, universities and affiliated research centers, technology companies, and a multinational law firm. While not all of the authors currently work at institutes of higher education, they bring a diverse collection of scholarly perspectives on frontier AI governance to the proposal.

This proposal is notable not only because of its collection of authors but also because of its extensive scope and policy suggestions, covering various mechanisms to develop, implement, and comply with AI standards. The proposal is also significant because it notes preconditions to implementing its recommendations, such as providing auditors with adequate resources and time to conduct sufficiently rigorous work, as well as pointing out uncertain areas of analysis, such as whether scaling of AI resources will continue along the same trajectory in the future.

Table 4: Key Components of Academia’s Frontier AI Regulation: Managing Emerging Risks to Public Safety

<table>
<tr><th colspan="3"><b>Why govern:</b> Government should create regulatory solutions that effectively minimize risks from frontier AI, which can possess dangerous capabilities that pose severe risks to public safety</th></tr>
<tr><th>What to govern:</th><th>Who governs:</th><th>How to govern:</th></tr>
<tr><td rowspan="4">Frontier AI safety standards development</td><td>AI developers</td><td>● Pilot and, if necessary, comply with safety standards during AI development and deployment, which may include:
○ Assessing models for capabilities and controllability
○ Repeatedly conducting risk assessments and updating deployment guardrails or rolling back models as needed
○ Tracking user behavior and cutting-edge research
○ Reporting AI incidents
○ Monitoring AI impacts</td></tr>
<tr><td>AI ethics and safety experts, AI researchers, academics, and consumer representatives</td><td>● Convene to build safety standards.</td></tr>
<tr><td>Governments</td><td>● Pioneer the development of test, evaluation, validation, or verification methods by updating procurement requirements.
● Fund research on emerging frontier AI risks and offer compute resources.
● Convene stakeholders.
● Provide guidance on frontier AI.</td></tr>
<tr><td>Third party auditors and red teamers</td><td>● Apply external scrutiny to AI models by eliciting capabilities.</td></tr>
<tr><td>Regulatory visibility</td><td>Regulators</td><td>● Develop a framework that facilities voluntary disclosure of information about frontier AI by companies.
● Mandate disclosures and impose reporting requirements on companies.</td></tr>
</table>

		● Audit companies against safety and risk management frameworks. ● Establish whistleblower regimes that protect individuals who disclose safety-critical information to relevant government authorities.
Standards compliance	Governments	● Encourage voluntary self-regulation and certification by implementing risk governance frameworks internally and encouraging the creation of a third-party compliance ecosystem. ● Mandate compliance with safety standards. ● Empower a supervisory authority to take enforcement measures. ● Require government license to widely develop and deploy a frontier AI model that poses risks to public safety above a certain threshold of severity. ● Provide subsidies and support to limit licensing compliance costs for small organizations.

Source: CSET.

The proposal identifies problems that exacerbate the challenge of regulating frontier AI systems, including the inability to anticipate every dangerous capability that a system may possess, and describes three pillars for regulating frontier AI. These include mechanisms for the development of frontier AI safety standards, such as sustained multi-stakeholder processes; mechanisms to give regulators visibility into frontier AI development, such as information disclosures; and mechanisms to ensure compliance with safety standards, such as licenses for frontier AI. Finally, the proposal suggests concrete measures that developers can take to ensure that their models are safe, including external scrutiny of models, deployment protocols that are proportionate to risk, and monitoring of model capabilities.

State Government: California's Safe and Secure Innovation for Frontier Artificial Intelligence Models Act

Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, or SB-1047, was introduced to the California (CA) legislature in February of 2024 by Senator Scott Wiener and co-authored by Senators Richard Roth, Susan Rubio, and Henry Stern.[10]

The bill was one of the first frontier AI governance proposals to be officially introduced to a state legislature and came close to becoming law, though it was eventually vetoed by Governor Gavin Newsom in September of 2024.[11] The bill was iterated upon over the course of its life as civil society advocates, AI developers, the tech industry, and other interested parties debated its merits and weaknesses. Table 5 breaks down the components of the bill prior to its veto.

This proposal was selected because it greatly influenced public discourse about AI governance and offers a template for other states to learn from. If enacted, SB-1047 could have significantly affected the future of frontier AI by placing requirements on the top U.S. AI companies, which happen to reside in California. Furthermore, pieces of the high-profile bill could inform other state AI governance efforts in the future given its near-passage.

Table 5: Key Components of SB-1047

Why govern: The California state government should mitigate risks from frontier AI while promoting innovation, competition, benefit-sharing, and equity

What to govern:	Who governs:	How to govern:
Frontier model development	Developers of AI models	● Ensure that models are capable of being fully shut down. ● Create and implement a safety and security protocol. ● Provide and retain a copy of the protocols and send it to the CA Attorney General. ● Retain a third-party auditor to evaluate compliance. ● Send the audit report to the CA Attorney General. ● Provide reasonable internal process for employee disclosure of concerns.
Frontier model deployment and use	Developers of AI models	● Do not use or provide access to frontier AI models if there is an unreasonable risk of critical harm. ● Ensure model harms are attributable before using or making a model available. ● Report any AI incidents that increase the risk of critical harm, such as inadvertent release of model weights, to the CA Attorney General.
Developers of AI models	Third-party auditors	● Audit developers' compliance with SB-1047, in keeping with regulations of the Government Operations Agency. ● Produce an audit report.
Thresholds and regulations determining which models are covered by SB-1047	Government Operations Agency	● Beginning in 2027, issue regulations updating SB-1047's definition of "covered model" to reflect scientific developments and international standards. ● Establish binding auditing requirements. ● Issue guidance for preventing unreasonable risk. ● House new Board of Frontier Models, which approves these guidances, auditing requirements, and regulations.
Compute access	Operators of computing clusters	● Govern customer use of compute and evaluate their usage by: ○ Identifying customers, their payment method, and IP address

		○ Assessing whether the compute is being used to train covered models ○ Enacting a full shutdown of compute access if needed ● Provide records of compute usage to the Attorney General.
Enforcement of the bill	CA Attorney General	● Bring civil action to enforce the law. ● Enforce whistleblower protections alongside the Labor Commissioner.
Promotion of AI safety research, innovation, and fair access to compute resources	Government Operations Agency	● Host consortium which creates CalCompute, a public cloud computing cluster.

Source: CSET.

Under SB-1047, developers of frontier AI models must ensure that their models are safe and face civil liability if they do not demonstrate due diligence. The bill applies to models that are above a certain compute threshold (the bill refers to these as “covered models”) and asks developers to evaluate whether their model is capable of causing “critical harm,” which it defines as:

- Chemical, Biological, Radiological, or Nuclear (CBRN) risks that could cause mass casualties,

- Cyberattacks on critical infrastructure causing mass casualties or $500,000,000 in damages,

- Criminal acts by a model acting without human oversight to cause mass casualties or $500,000,000 in damages, or

- Other grave harms to public safety that are comparably severe.

If there is a risk of a model causing or enabling critical harm, developers are prohibited from using the model or making it publicly available. Furthermore, developers of frontier AI models must be able to fully shut down their model, implement safety protocols, conduct third-party audits, and submit annual certifications to the California government proving their compliance with the law. The bill directs compute operators to evaluate customer activity to determine whether compute is being used to train covered models, and instructs the California government to create a new compute resource to facilitate AI safety research and offer compute access to under-resourced research communities. Finally, the bill contains whistleblower protections for employees reporting on their company's violations of the bill.

Federal Government: Framework to Mitigate AI-Enabled Extreme Risks

The congressional "Framework to Mitigate AI-Enabled Extreme Risks," co-authored by Senators Mitt Romney, Jack Reed, Jerry Moran, and Angus King and unveiled in April of 2024, is much shorter and less specific than the other proposals.[12] Nevertheless, we selected the proposal because it offers insight into how several members of Congress are thinking about frontier AI and represents one of the first congressional frameworks to deal exclusively with extreme risks posed by advanced models.[13] The proposal emphasizes CBRN risks, U.S. national security, innovation, and a new oversight entity that could take on regulatory responsibilities. The original framework drafters, as well as Senator Maggie Hassan, have since turned the framework into a bill, the Preserving American Dominance in AI Act, though our analysis still focuses on the initial framework.[14]

Table 6: Key Components of Framework to Mitigate AI-Enabled Extreme Risks

Why govern: The federal government should reduce CBRN risks from frontier AI and protect U.S. national security without harming innovation		
What to govern:	**Who governs:**	**How to govern:**
Federal authority to govern frontier AI	Congress	- Give authority to new oversight entity, which can be a new interagency coordinating body, a preexisting federal agency, or a new agency.
Frontier AI hardware and compute	Oversight entity	- Receive reports from entities selling and buying compute. - Study and report to Congress on emerging challenges to ensure the framework's provisions remain appropriate as tech advances.
Developers of AI models	Compute providers	- Report large acquisitions or usage of compute to the oversight entity. - Screen customers, paying close attention to foreign persons.
Frontier AI development	Oversight entity	- Receive notification from model developers building frontier AI.
Compute access	Developers of frontier AI systems	- Notify the oversight entity when developing a frontier model and prior to initiating training runs. - Incorporate safeguards against CBRN risk. - Adhere to cybersecurity standards to prevent model leaks or theft. - May be required to report to the oversight entity on efforts to mitigate CBRN risk and implement cybersecurity standards.
Frontier AI deployment	Oversight entity	- Provide tiered licenses to frontier model developers based on the results of evaluations assessing CBRN risk.

Source: CSET.

The congressional framework emphasizes cybersecurity by advocating for protections against the theft of frontier models as well as the cybersecurity risks that models pose. Like SB-1047, the proposal directs compute providers to screen their customers and includes a compute threshold as a method for defining frontier AI. The proposal does not, however, specify any consequences that developers or compute providers would face if they were to violate its provisions.

Analysis

Deriving Assumptions

We examine each proposal's underlying assumptions, or conditions that must be true in order for the proposal to be effective. While our approach bears some similarities to methodologies such as assumption-based planning, in which assumptions are deliberately identified and used to generate a plan or inform an organization's planning process, we focus specifically on assumptions contained within the proposals themselves.[15]

The assumptions we identify in the Analysis section are responses to questions that we derive from interactions between a proposal's key components. These questions are:

1. **Which risks are important to mitigate and who should have primary oversight of frontier AI?** In other words, which risks should be addressed and which actors should be charged with overseeing frontier AI? These questions were derived from the "why govern" key component of proposals.

2. **Who is delegated tasks and able to play a role?** Simply, does the proposal identify actors who can actually and effectively use the governance tools and mechanisms outlined in the proposal? These actors must exist, now or in the future, and have the capacity to leverage tools and mechanisms. This question was derived by examining the relationship between the "who governs" and "how to govern" key components.

3. **Would the proposed mechanisms or tools actually achieve the proposal's objectives?** In other words, would the proposal's mechanisms or tools accomplish the proposal's objectives and align with the proposal's focus areas? This question was derived by examining the relationships between the "why govern" (proposal's objectives), "what to govern" (broad areas being governed), and "how to govern" (specific mechanisms or tools) key components. Tables 9, 10, and 11 phrase this as:

    - Which techniques or categories of techniques are effective?

    - Which processes or frameworks are necessary?

    - Which information or actions are useful?

Some of these assumptions are shared between proposals, and some are unique. The following section examines assumptions that are shared across multiple proposals. Understanding shared assumptions can help policymakers identify opportunities for consensus-building among different stakeholders and surface ways to facilitate the success of multiple proposals. We primarily analyze each shared assumption in isolation but discuss how interdependencies among sets of assumptions can impact the effectiveness of governance proposals in Appendix A. On the other hand, identifying unique assumptions can highlight differences between proposals as well as areas where consensus-building may be challenging.

Shared Assumptions

Shared assumptions are held across at least two proposals due to the limited number of case studies we consider in this report. For policymakers considering a wider variety of proposals, it may be useful to set a higher threshold for what constitutes a shared assumption. Colored cells in the tables below indicate that a proposal includes the corresponding assumption. Blank cells in the tables do not necessarily mean that the proposal disagrees with the assumption but rather that the proposal does not explicitly address the assumption.

Table 7. Shared Assumptions About Which Risks Are Important to Mitigate and Who Should Have Primary Oversight of Frontier AI

Which risks are important to mitigate and who should have primary oversight of frontier AI?	OpenAI Proposal	Zero Trust AI Governance	Managing Emerging Risks to Public Safety	SB-1047	Framework to Mitigate AI-Enabled Extreme Risks
Frontier AI could present severe catastrophic risks that are important to mitigate.	🟩		🟩	🟩	🟩
The majority of oversight of frontier AI should fall to government actors.*		🟩	🟩	🟩	🟩
The majority of oversight of frontier AI should fall to industry actors.	🟩				

Source: CSET.

What Risks Are Important to Mitigate and Who Should Have Primary Oversight of Frontier AI?

Table 7 summarizes how proposals address questions regarding risk and responsibility. The majority of proposals assert that frontier AI could present severe catastrophic risks, such as CBRN risk. This likely reflects concerns about AI systems potentially enhancing the abilities of malicious actors. While proposals name a variety of actors responsible for addressing frontier AI risks (including third-party auditors, federal regulators, Congress, experts in fields other than AI, and computing resource providers), most proposals task government actors with the primary oversight role of frontier AI systems.

Determining who bears primary oversight responsibilities is not always a straightforward task. For example, SB-1047 holds industry actors civilly liable for

* By "oversight," we mean maintaining shared infrastructure, overseeing proposals' regulatory efforts, and enforcing these efforts in the face of noncompliance.

damages caused by their products unless they implement safety and security protocols and comply with the law, but it directs the state government to update the definition of a “covered model” and receive reports from companies about their protocols. In this case, both industry and government have some degree of oversight responsibility. Nevertheless, we determine that the majority of oversight responsibility rests on the state government because the law directs the California government to maintain public infrastructure, provide guidance on complying with the law, and ultimately enforce the law. We made similar determinations for the federal government, academic, and civil society proposals.

On the other hand, the OpenAI proposal centers entirely on actions that OpenAI can take to mitigate risks, suggesting that the authors of the proposal believed that such internal governance would help significantly mitigate catastrophic risks. Another interpretation might be that OpenAI’s proposal was intended to create a model for other firms to follow, encouraging behaviors that would help reduce risk across the AI industry. Although the assumption that industry actors should have primary oversight over frontier AI systems is unique to OpenAI, it is included in the table to highlight its contrast with the previous assumptions.

Who Is Delegated Tasks and Able to Play a Role?

Another class of shared assumptions centers on the parties who are expected to execute the proposals. All of the proposals delegate governance responsibilities to key actors, although the scope of the responsibilities and the types of actors vary by proposal.

Table 8. Shared Assumptions About Who is Delegated Tasks and Able to Play a Role

Who is delegated tasks and able to play a role?	OpenAI Proposal	Zero Trust AI Governance	Managing Emerging Risks to Public Safety	SB-1047	Framework to Mitigate AI-Enabled Extreme Risks
Government has sufficient talent and capacity to accomplish technical tasks such as standard-setting.		🟩	🟩	🟩	🟩
A variety of experts, including AI ethics and safety experts, AI researchers, and consumer representatives, have the ability to participate in standard-setting.		🟩	🟩		
A third-party auditing ecosystem exists and has the capacity and access to conduct audits of AI systems.	🟩	🟩	🟩	🟩	
Developers are able to operationalize high-level guidance from authorities.		🟩	🟩	🟩	🟩

Source: CSET.

The assumptions in Table 8 relate to the abilities of key actors to carry out responsibilities in the proposal.[*] For instance, four out of the five proposals hinge on the assumption that government actors are capable of accomplishing technical tasks related to AI governance. Furthermore, four out of the five proposals assume that a third-party auditing ecosystem exists and is capable of conducting audits of frontier AI systems and that developers are able to operationalize high-level guidance from authorities. If these underlying assumptions are not true, then implementation of most proposals will require provisions to upskill key actors, attract and retain talent, or clarify high-level guidance.

The abilities of key actors to carry out their responsibilities are shaped by factors including political will, institutional support, and shared conceptual understanding of AI risk mitigations. Political will is needed to govern companies that may resist regulation and mobilize resources to lobby against it. Political will is also important for government actors because they have limited resources and there is an opportunity cost associated with pursuing certain priorities at the expense of other projects. For example, research has shown that the United States is facing an accelerated STEM talent crisis.[16] Without the political will to attract and retain tech talent, it is unlikely that the U.S. government will have adequate resources to grow the AI workforce and effectively govern AI. Institutional support in the form of financial resources and logistical services can also facilitate diverse groups' contributions to AI safety standards and access to frontier AI systems. Finally, proposals can be streamlined if assumptions about shared conceptual understanding hold true. For example, proposals do not have to elaborate on what AI risk management entails if relevant stakeholders already understand how to operationalize high-level risk management guidance. However, such assumptions may lead to implementation problems if relevant stakeholders do not agree on ways to translate AI risk management into practice.

Would the Proposed Mechanisms or Tools Actually Achieve the Proposal's Objectives?

Finally, we examine assumptions related to whether the mechanisms and tools recommended by proposals align with the proposals' focus areas and accomplish the

[*] The assumptions in Tables 8 through 11 are framed in the present tense, although the lack of task sequencing in proposals makes it unclear whether the proposals' authors believed that assumptions held at the time of writing or would hold in the future.

stated goals of proposals. These assumptions are wide-ranging and diverse, and fall into three general categories:

1. Which techniques or categories of techniques are effective
2. Which processes or frameworks are necessary
3. Which information or actions are useful

These assumptions tend to surface material prerequisites that must be in place to achieve the proposals' AI governance objectives. For example, if a proposal relies on incident reporting as a governance mechanism, then frameworks for reporting incidents in a consistent and understandable way are needed for the proposal to work.

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.

Most assumptions about techniques that are effective at mitigating AI risks and harms are not widely shared across proposals (Table 9). The assumptions that preventing model leakage or theft, watermarking or implementing content provenance techniques, attributing harms from AI, and monitoring customer use of compute are effective are shared across two or three (but not always the same two or three) proposals. The uneven distribution of assumptions reflects the general lack of consensus regarding which techniques are most effective at mitigating AI harms and risks. Watermarking and content provenance techniques are some of the most readily implementable techniques listed in the table because tools for these techniques are already available.[17] However, researchers have debated the robustness of watermarking and content provenance techniques in practice.[18] The other mechanisms in Table 9 would require sophisticated monitoring infrastructure, cybersecurity techniques, or root cause analyses — none of which are fully developed for AI.

The only assumption that is shared across four proposals involves identifying and tracking frontier AI risks before they become harms. This assumption may reflect shared concerns about addressing prospective or speculative risks from frontier AI. Zero Trust AI Governance, the one proposal that does not include this assumption, focuses on classes of harm that the other proposals do not, such as consumer protection and civil rights violations.

Table 10. Shared Assumptions About Which Processes or Frameworks Are Necessary

Which processes or frameworks are necessary?	OpenAI Proposal	Zero Trust AI Governance	Managing Emerging Risks to Public Safety	SB-1047	Framework to Mitigate AI-Enabled Extreme Risks
Safety and risk management frameworks or standards	🟩	🟩	🟩	🟩	🟩
Compliance mechanisms		🟩	🟩	🟩	🟩
Imposition of liability		🟩		🟩	
Information sharing between developers, users, and government actors	🟩	🟩	🟩	🟩	🟩
Mechanisms to monitor AI impacts	🟩	🟩	🟩	🟩	
Incident reporting frameworks		🟩	🟩	🟩	
Whistleblower protections for employees			🟩	🟩	

Source: CSET.

Compared to Table 9 (effectiveness of techniques), Table 10 reflects a greater degree of consensus in the assumptions shared among proposals, suggesting that establishing these processes and frameworks could benefit multiple governance proposals. Two types in particular are shared across all proposals: safety and risk management frameworks or standards, and information sharing between AI developers, users, and government actors. Two others are shared by four of the five proposals: mechanisms to ensure compliance with each proposal's rules and processes for AI impact monitoring. Furthermore, three proposals assume that AI incident reporting frameworks are necessary (the OpenAI proposal mentions bug bounties but not incident reporting). These frameworks and processes can help enforce policies, provide important data about whether policies are actually effective at mitigating risks from AI systems, or establish accountability structures. SB-1047 in particular includes several measures to promote accountability, including whistleblower protections for employees reporting inappropriate conduct, a framework to track AI incidents (which can help minimize the likelihood that AI incidents happen again), and imposition of civil liability on developers of covered models.

Both Table 8 (availability and capacity of key actors) and Table 10 highlight AI-enabling talent, processes, and frameworks as important precursors for AI governance. They also highlight the nontrivial work that is required to implement these governance proposals, given that several of the referenced frameworks and sources of talent are not fully developed. For example, incident reporting frameworks are not standardized nor have they been widely adopted. Furthermore, research has demonstrated barriers to a robust third-party auditing ecosystem that range from lack of auditor independence to poorly scoped audit standards, suggesting that the availability and capacity of key actors in proposals may be limited.[19]

Table 11. Shared Assumptions About Which Information or Actions Are Useful

Which information or actions are useful?	OpenAI Proposal	Zero Trust AI Governance	Managing Emerging Risks to Public Safety	SB-1047	Framework to Mitigate AI-Enabled Extreme Risks
Using compute thresholds as a proxy for performance and risk				🟩	🟩
Disclosures and reports from companies about AI development	🟩	🟩	🟩	🟩	🟩
Customer vetting and screening for compute access				🟩	🟩
Reports from entities selling and buying compute				🟩	🟩
Licensing frontier AI systems			🟩		🟩
Providing broader access to compute resources			🟩	🟩	

Source: CSET.

Finally, Table 11 presents information and actions that proposals assume are useful for governing AI systems. All five proposals assume that disclosures from AI companies about AI development are necessary. This includes the OpenAI proposal, which does not specifically mention disclosing information to the government but does propose a reporting structure to inform other AI labs about model vulnerabilities post-release and suggests disclosing information through a system card.* Other assumptions, like the need for broader compute access, customer screening, or licensing regimes, are shared more disparately. Notably, the two government proposals—SB-1047 and the federal framework to mitigate extreme risks from AI—share assumptions regarding compute thresholds, customer screening for compute access, and reporting from compute providers. This is unsurprising given that placing limits on compute access and usage is commonly touted as an AI governance lever that governments can adjust fairly easily, as compute demands change over time.[20]

While all of the assumptions in Table 11 concern information or actions that are intended to help minimize risk, the final two—licensing frontier AI systems and expanding access to compute resources—hold true only if they strike a balance between mitigating AI risks and promoting AI innovation. A balanced licensing regime for frontier AI must protect national security without making it overly challenging for less established AI firms to obtain a license, and managers of compute access must weigh the benefits of expanding access—including more competition, benefit-sharing, and equity—with the risks from providing more organizations with key inputs to develop (and potentially misuse) frontier AI.

Unique Assumptions

In addition to the assumptions shared across proposals, two proposals contain unique assumptions that address either the stage of the AI system lifecycle during which risks should be managed or the degree to which actors can control a model.

AI governance proposals emphasize different stages of the AI system lifecycle when prescribing how to manage AI risks. The AI system lifecycle can be roughly divided into two parts: development (during which companies and organizations are building a model but have not yet released it to the public) and deployment (the time after which a model has been released). The OpenAI proposal assumes that risks are best managed at the deployment stage, as it includes more risk mitigation provisions for the deployment stage than the development stage. This is in contrast to other governance proposals we analyzed. For example, the Zero Trust AI Governance proposal assumes

* A system card is documentation that explains how an AI system works.

that risks are best managed by actors across the entire AI system lifecycle, encompassing both development and deployment. Although most of the proposal focuses on the post-deployment stage, it highlights the importance of sound development practices and ethically sourced training data. The academic proposal also advocates for interventions that span the development, deployment, and post-deployment stages. The contrast of OpenAI's unique assumption with other proposals reveals different ways of thinking about AI risk management that map roughly onto the stages of the AI system lifecycle.

SB-1047 contains a unique assumption about the extent to which actors can control a model. More specifically, SB-1047 assumes that AI developers can completely shut down a model and that compute providers have the ability to completely cut off a customer's access to their computing resources. It is the only proposal of the five to assume the existence of "kill switch" mechanisms. Among the proposals that address computing power as an AI governance mechanism, it also assigns the most responsibility to compute providers in the case of an AI-related emergency. In SB-1047's case, these unique assumptions highlight the relative importance of computing power to the proposal's authors, and how AI-related policies can encompass governance of adjacent technologies.

Discussion

Our analysis of the assumptions that underpin five AI governance proposals reveals that policymakers can use assumptions to better understand the contours of AI governance debates and pursue avenues for collaboration, coalition building, or legislative action.

1) Policymakers should leverage proposals' assumptions to more precisely understand disagreements and shared views among stakeholders. Surfacing underlying assumptions provides a structured mechanism for navigating disagreement. This process can reveal that what appears to be disagreement among stakeholders may actually stem from different assessments of assumptions that can be empirically investigated or otherwise addressed. For example, the OpenAI proposal, SB-1047, and the federal framework share the assumption that preventing model leakage and theft is technically feasible, whereas the civil society and academic proposals do not address this assumption. It may be that the authors of the civil society and academic proposals do not view preventing model leakage and theft as integral to AI governance; alternatively, the authors may view model leakage and theft as an important and tractable problem but one that will not be solved for a long time. In this case, abstract disagreements among stakeholders reveal hypotheses that can be further examined.

Conversely, surfacing underlying assumptions can reveal when areas of agreement obscure greater complexity. In particular, policymakers should be wary of conceptually grouping stakeholders together based on shared risk assessments. Several proposals are motivated by a desire to prevent catastrophic risks related to frontier AI but focus on different mechanisms for mitigating those risks, suggesting that their recommendations are not easily interchangeable. For example, the academic and state government proposals both suggest that frontier AI could present severe catastrophic risks, but of the two, only SB-1047 mandates reports from entities buying and selling compute. Additionally, the academic proposal advocates for a licensing regime whereas SB-1047 does not. Licensing and reporting are two mechanisms that place responsibilities of varying difficulty levels on different stakeholders. A licensing regime would require the designation of an oversight entity and compliance from developers, whereas mandating reports from entities buying and selling compute would be a relatively lower lift and involve placing responsibilities on compute providers. In some cases, policymakers may wish to harmonize proposals' recommendations and create a more unified approach; in others, they may use proposals to gain an awareness of different positions on frontier AI governance and ensure that the policies they pursue are well-informed. Either way, policymakers will benefit from a more detailed understanding of the assumptions underlying AI governance proposals.

2) Policymakers can take action in an uncertain and rapidly changing environment by addressing common assumptions across proposals. Multiple proposals share underlying assumptions about actor capabilities, the effectiveness of techniques related to governing frontier AI, the existence and maturity of frontier AI processes and frameworks, and the utility of certain information and actions. These represent common prerequisites, conditions or tangible requirements that are necessary (but not always sufficient) to fulfill multiple proposals. However, these prerequisites do not always exist and may require support from policymakers to become a reality. For example, all five proposals require some form of information sharing between government actors, developers, and users. Policymakers can make each of these proposals easier to implement by funding research and development of information sharing channels and integrating effective information sharing approaches into current and future policies.

Alternatively, shared prerequisites may not always indicate a need to invest in a specific resource but instead represent avenues for more investigation, productive collaboration, or coalition building. For example, many proposals reference common actors such as third-party auditors, experts in fields other than AI, and computing resource providers. Policymakers who direct any one of these actors to carry out policies may overburden them, especially if policymakers intend to implement multiple proposals. To reduce the risk of overburdening these actors, policymakers could instead investigate how to proportionately distribute responsibilities among common actors, foster partnerships between them, or incentivize new actors to join the AI community to benefit multiple AI governance proposals.

Conclusion

In this paper, we conduct case studies of five AI governance proposals and introduce an analytic approach for identifying assumptions in policy proposals. We identify two key takeaways: 1) Policymakers should leverage proposals' assumptions to more precisely understand disagreements and shared views among stakeholders and 2) policymakers can take action in an uncertain and rapidly changing environment by addressing common assumptions across proposals.

Forthcoming AI governance proposals will likely recommend different actions than the proposals we examine due to societal changes and the evolution of AI technology. However, some of the assumptions that underpin proposals may persist over time. Although the assumptions in our analysis were derived from proposals that are a couple of years old, they are still relevant at the time of writing. For example, the need for actionable risk management frameworks and effective content provenance techniques are still top of mind for policymakers and technologists alike. We expect that newer proposals will rely on some of the same assumptions as older proposals, even if their details differ.

We advise policymakers to identify the assumptions that underpin AI governance proposals, which may have more staying power than the details of proposals themselves. Policymakers can leverage assumption-based analysis to move beyond rhetorical debates and contribute to more productive policy discussions. By addressing these assumptions, policymakers can help shape the AI ecosystem now to prepare the United States for a range of possible AI futures.

Authors

Mina Narayanan is a research analyst at CSET, where she works on AI governance and safety.

Jessica Ji is a senior research analyst at CSET, where she works on the CyberAI Project.

Vikram Venkatram is a research analyst at CSET, where he focuses on emerging issues in biotechnology.

Ngor Luong contributed to this research while she was a senior research analyst at CSET. She is currently a senior policy analyst at the U.S.-China Economic and Security Review Commission.

Acknowledgments

We would like to thank Owen Daniels, Steph Batalis, Zachary Arnold, Jack Corrigan, Mia Hoffmann, Kyle Miller, Helen Toner, Nicklas Lundblad, Joseph Keller, Thomas Woodside, Dale Brauner, and Kendrea Beers for their thoughtful feedback and for holding our writing to a high standard.

Document Identifier: doi: 10.51593/20240005

Appendix A: Assumption Interdependencies

Many of the assumptions that we derive can be grouped into broader categories to better illustrate their interdependencies. These categories include factual assumptions about the state of the world, capacity assumptions about the expertise of actors and the resources they are afforded, mechanism assumptions about the feasibility and effectiveness of different techniques, and normative assumptions about values and how the world should be. These sets of assumptions may interact, enable, or constrain one another in ways that impact the effectiveness of governance proposals. For instance, assumptions about the effectiveness of mechanisms build upon factual assumptions that the problems that mechanisms target are real and pressing, and normative assumptions about what should be done may be constrained by available resources and capacity. In the context of our report, a normative assumption that the majority of oversight for frontier AI systems should fall to government actors is contingent on a capacity assumption that the government has adequate resources to provide oversight.

These interdependencies highlight the importance of identifying sets of assumptions that must collectively hold true for governance proposals to be effective. Policymakers can use multi-dimensional analysis that accounts for different categories of assumptions to better anticipate potential points of failure and more effectively design governance mechanisms that are resilient under a variety of scenarios. However, this type of analysis is beyond the scope of this report.

Endnotes

[1] Melissa Heikkilä, “Dutch Scandal Serves as a Warning for Europe Over Risks of Using Algorithms,” Politico, March 29, 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/; Julie Jargon, “Fake Nudes of Real Students Cause an Uproar at a New Jersey High School,” *The Wall Street Journal*, November 2, 2023, https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb?mod=tech_lead_story.

[2] Kashmir Hill, “She Is in Love With ChatGPT,” *The New York Times*, January 17, 2025, https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html; Cecily Mauran, “120 Court Cases Have Been Caught With AI Hallucinations, According to New Database,” Mashable, May 27, 2025, https://mashable.com/article/over-120-court-cases-caught-ai-hallucinations-new-database.

[3] Helen Toner, Jessica Ji, John Bansemer et al., "Skating to Where the Puck Is Going" (Center for Security and Emerging Technology, October 2023), https://cset.georgetown.edu/publication/skating-to-where-the-puck-is-going/.

[4] "OpenAI’s Approach to Frontier Risk," OpenAI, October 26, 2023, https://openai.com/global-affairs/our-approach-to-frontier-risk/.

[5] Department for Science, Innovation and Technology, “A Pro-Innovation Approach to AI Regulation: Government Response” (London: Department for Science, Innovation and Technology, February 6, 2024), https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response.

[6] “OpenAI Red Teaming Network,” OpenAI, September 19, 2023, https://openai.com/index/red-teaming-network/; Will Knight, “OpenAI’s Long-Term AI Risk Team Has Disbanded,” *Wired*, May 17, 2024, https://www.wired.com/story/openai-superalignment-team-disbanded/.

[7] AI Now Institute, Accountable Tech, & Electronic Privacy Information Center (EPIC), “Zero Trust AI Governance” (August 2023), https://ainowinstitute.org/wp-content/uploads/2023/08/Zero-Trust-AI-Governance.pdf.

[8] Ryan Terry, “Zero Trust Security Explained: Principles of the Zero Trust Model,” CrowdStrike, March 13, 2025, https://www.crowdstrike.com/en-us/cybersecurity-101/zero-trust-security/.

[9] Markus Anderljung, Joslyn Barnhart, Anton Korinek et al., “Frontier AI Regulation: Managing Emerging Risks to Public Safety,” arXiv preprint arXiv:2307.03718 (2023), https://arxiv.org/abs/2307.03718.

[10] Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, SB-1047, California Legislature (2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047.

[11] “SB-1047 Veto Message,” Office of the Governor of California, September 29, 2024, https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf.

[12] "Framework for Mitigating Extreme AI Risks," Office of U.S. Senator Mitt Romney, April 16, 2024, https://web.archive.org/web/20241223104639/https://www.romney.senate.gov/wp-content/uploads/2024/04/AI-Framework_2pager.pdf#expand.

[13] "King, Colleagues Unveil Bipartisan Framework to Identify, Minimize Artificial Intelligence Risks," Office of U.S. Senator Angus King, April 16, 2024, https://www.king.senate.gov/newsroom/press-releases/king-colleagues-unveil-bipartisan-framework-to-identify-minimize-artificial-intelligence-risks.

[14] Preserving American Dominance in Artificial Intelligence Act of 2024, S. 5616, 118th Cong. (2024), https://www.congress.gov/bill/118th-congress/senate-bill/5616/text; "King, Colleagues Introduce Bipartisan 'Preserving American Dominance in AI Act,'" Office of U.S. Senator Angus King, December 20, 2024, https://www.king.senate.gov/newsroom/press-releases/king-colleagues-introduce-bipartisan-preserving-american-dominance-in-ai-act.

[15] James A. Dewar, *Assumption-Based Planning: A Tool for Reducing Avoidable Surprises* (New York: Cambridge University Press, 2002), https://www.cambridge.org/core/books/assumptionbased-planning/CEB920081B04403472F14DD77E66E2C9; Sammy Davis-Mendelow, Jorge A. Baier, and Sheila A. McIlraith, "Assumption-Based Planning: Generating Plans and Explanations Under Incomplete Knowledge," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* 27, no. 1 (June 2013): 1075–1081, https://aaai.org/papers/8687-assumption-based-planning-generating-plans-and-explanations-under-incomplete-knowledge/.

[16] National Science Board, *Talent Is the Treasure* (Washington, DC: National Science Foundation, 2024), https://www.nsf.gov/nsb/publications/2024/2024_policy_brief.pdf.

[17] "C2PA: An Open Technical Standard," Coalition for Content Provenance and Authenticity (C2PA), accessed May 19, 2025, https://c2pa.org/; "AI Watermarking 101: Tools and Techniques," Hugging Face Blog, August 17, 2023, https://huggingface.co/blog/watermarking; Meta AI, "Stable Signature: A New Method for Watermarking Images Created by Open Source Generative AI," Meta, November 7, 2023, https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/; "SynthID: Watermarking and Identifying AI-Generated Content," Google AI, August 29, 2023, https://ai.google.dev/responsible/docs/safeguards/synthid.

[18] Nikola Jovanović, Robin Staab, and Martin Vechev, "Watermark Stealing in Large Language Models," arXiv preprint arXiv:2402.19361 (2024), https://arxiv.org/abs/2402.19361.

[19] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho, "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, July 2022, https://doi.org/10.1145/3514094.3534181.

[20] Exec. Order No. 14110, 88 FR 75191 (2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence; "Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," European Parliament and Council of the European Union, July 12, 2024, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.