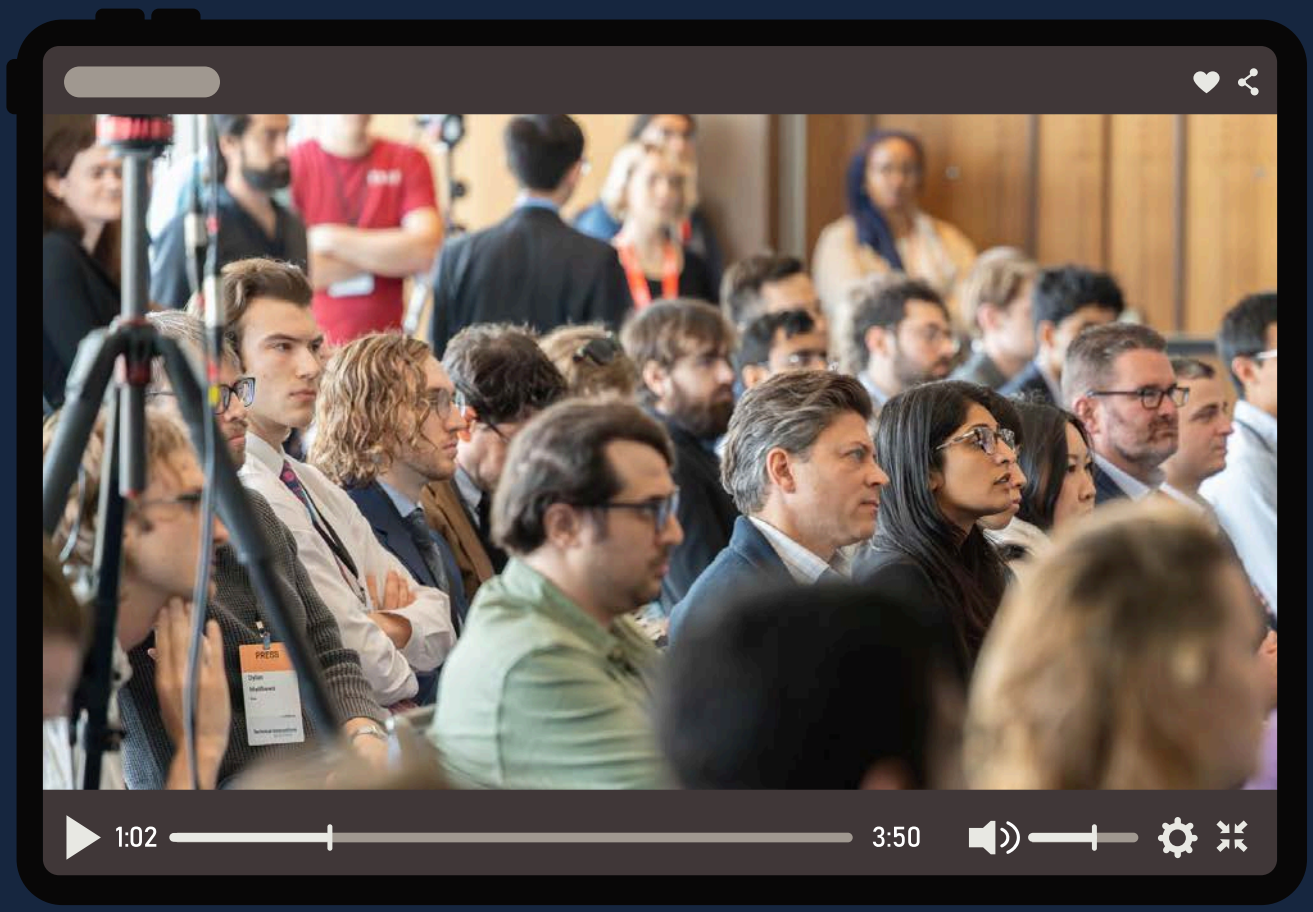
Technical Innovations for AI Policy

Challenges in AI Governance

Key insights from 15 expert talks



FAR.AI

Frontier AI Safety Policy

Expectant father Alex Bores draws on his wife's pregnancy to illustrate the rapid pace at which AI policy evolves. Seven months ago Anthropic called for comprehensive safeguards within 18 months; since then a Newsom commission has warned of the closing policy window for AI policy action, and Palisade documented OpenAI's o3 model refusing to follow shutdown commands. States like Colorado, Vermont, Texas, and California have sketched AI safety bills, with New York's RAISE Act is closest to passage. RAISE would apply to any lab spending $100 million or more on training, requiring a public safety plan, third-party audit, critical-incident disclosure, and strong whistle-blower protection. Alex ends by calling on the audience to reach out to their representatives and take direct action in the coming weeks while the narrow policy window remains open.

Alex Bores

New York State Assembly

Unresolved Debates on the Future of AI

Helen Toner's dissects three technical disagreements at the heart of AI policy. First: how far can the current paradigm go? GPT-style models continue advancing through reasoning training and multimodality, but face limits from hallucinations, reliability gaps, memory constraints, and diminishing training returns. Second: how much can AI improve AI? Google has already used Gemini to speed model training by 1%, saving millions of dollars, though bugs, lack of research judgment, and physical constraints may slow progress. Third: will future AI basically remain tools, or become something else? Today's models behave like conventional tools, but unlike traditional technology, AI is "grown, not built," shows emerging awareness, and faces market pressure toward autonomy—potentially resembling self-sustaining systems more than previous technologies.

Helen Toner

Center for Security & Emerging Technology

DoD & AGI Preparedness

Mark Beall argues that AGI represents the defining challenge of our era, surpassing nuclear threats and great-power competition. He identifies a dangerous cultural divide between Silicon Valley and Washington that has left legislators unprepared for AGI's arrival and now poses a greater threat than any foreign adversary. With potentially only three years to act, he proposes wedding "acceleration and altruism" through a three-pronged strategy: protect (export controls, deterrence), promote (defense AI adoption), and prepare (government-industry transparency, China diplomacy). He warns that without rapid cooperation—including technologists joining government and drastic measures like nationalization if necessary—America risks losing AGI leadership to authoritarian control.

Mark Beall

The AI Policy Network

13 (+1) Ways of Looking at AI

Brad Carson set out 14 of the unanswered questions that shape his thinking on AI policy. After examining public support for AI infrastructure, incentives for AI-driven bioweapons, Chinese semiconductor capabilities, and democracy's future amid AI automation, Carson singles out three areas to offer his preliminary views. First, the energy challenge: U.S. AI data centers may soon demand 500 TWh, equivalent to adding 75 Three Mile Island nuclear reactors. Second, AI's military implications: While current algorithms enhance intelligence capabilities, edge cases and unexpected chaos may limit their use in critical warfare domains. Third, a question of focus: policymakers have fixated on generative AI, yet recommender and predictive systems already govern credit, policing, and information consumption—often performing no better than human judgment—while quietly eroding civic foundations necessary for future regulation. Carson concludes by asking his audience to help refine these arguments, noting that definitive answers remain elusive.

Brad Carson

Americans for Responsible Innovation

Regulation of

Frontier Models, Confidential Computing & AI Alignment

Congressman Bill Foster focuses on three congressional priorities for AI governance. He proposes implementing secure digital IDs for Americans, preventing AI chip smuggling through location-proving circuitry and time-limited licensing, and establishing international cooperation for GPU licensing. Foster details the technical aspects of location verification using trusted sentinel modules and cryptographic challenges. He emphasizes the importance of hardware-enforced mechanisms to ensure compliance and prevent unauthorized use of high-end AI chips. Foster argues that this approach could foster meaningful discussions on AI safety by centering on the conditions for obtaining GPU operation licenses.

Bill Foster
U.S. House of Representatives

Day 1
Opening Remarks

Adam Gleave opened the Technical Innovations for AI Policy Conference by challenging the notion that AI policy must choose between innovation and safety. He argued that technical solutions could enable both progress and trustworthy development. Gleave cited examples like air pollution control and differential privacy as precedents for solving complex policy issues through innovation. He highlighted two pressing needs: secure third-party model evaluation for CBRN risk assessment and on-chip mechanisms to monitor AI chip usage for export control. He stressed the importance of collaboration between policymakers and technical experts from various sectors to develop and implement effective AI governance mechanisms.

Adam Gleave
FAR.AI

Making Sense of the AI Auditing Ecosystem

Miranda Bogen argues that vague terminology enables "checkbox compliance" that appears rigorous but ignores real risks. Her team categorized hundreds of assessment methods using two dimensions: scope (from broad exploratory red-teaming to narrow metric testing) and independence (from internal reviews to adversarial third-party audits). Each approach trades access for credibility. Understanding these trade-offs helps clarify what any proposal can actually achieve, reveals coverage gaps, and shows why no single method works alone. Effective governance requires defining clear goals, choosing the right mix of methods, setting specific metrics, funding independent oversight, and building feedback loops from development through deployment.

Miranda Bogen
Center for Democracy & Technology

AI Control:
Addressing Risks from Agentic Internal Deployments

Mary Phuong presents AI control as the second defense layer against misalignment. Since reliable alignment guarantees appear unlikely, AI control focuses on safe deployment even assuming all models are actively misaligned. Phuong projects that within a few years, AI labs will deploy AI agents that run autonomously for multiple days with access to sensitive internal systems—and in numbers too large for meaningful human oversight. Despite these challenges, she argues that AI control can deliver robust safety through automated monitoring, escalation processes, system design, and control evaluations. While Phuong sees AI control as our most effective near-term approach for mitigating misalignment risks, she cautions it doesn't address the fundamental challenge of preventing misaligned superintelligent AI systems.

Mary Phuong
Google DeepMind

Day 2
Opening Remarks

Lennart Heim identifies three key fronts of progress in technical AI governance: understanding frontier AI developments, analyzing their trends to inform governance solutions, and developing technical mechanisms and standards to implement those solutions. He illustrates these contributions with examples from the research on AI's exponential compute and energy demands, analysis comparing US and Chinese capabilities in energy infrastructure and chip production, and the development of verification systems that enable us to "trust math over people" for international agreements and AI company commitments. While technical AI governance has grown from a niche specialty to a field with hundreds of practitioners, Heim contends that there remain too few "adults in the room," which stresses the need for people who understand the technical fundamentals, conduct rigorous analysis, and find practical pathways for implementation within existing systems.

Lennart Heim
RAND

Overview of Technical AI Governance

Ben Bucknall defines technical AI governance as using technical analysis to identify governance needs and inform policy decisions. He highlights how policy aspirations often clash with technical reality, citing watermarking as an example where policymakers promoted uncertain solutions. Bucknall's framework categorizes governance by targets (data, compute, models, deployment) and capacities (assessment, verification, operationalization, etc.). The Foundation Model Transparency Index shows developers report well on capabilities but poorly on impacts. He emphasizes that effective governance requires bidirectional communication between policymakers who set goals and technical experts who determine feasible implementation.

Ben Bucknall
University of Oxford

AI in the National Security & Defense

Steve Kelly examines both the opportunities and risks that AI presents for national security. He addresses three key questions: AI's value in national security and defense, its potential role in geopolitical stability, and whether AI itself could become a national security threat. Kelly highlights the value of AI tools for intelligence gathering, military planning, and weapons systems; notes how China might rush to integrate AI to compensate for its lack of combat experience, and how integrating AI in our systems might lead to overreliance on it and a decline in our capacity to reason and make decisions of our own. Kelly concludes by emphasizing the need to plan for and avoid scenarios where humans lose meaningful control over AI systems.

Steve Kelly

Institute for Security & Technology

Policy-Oriented AI Evaluations

Kevin Wei outlines how to create AI capability assessments relevant to policymakers. Wei emphasizes three key points: focus on capabilities that will prompt policymaker action, choose the right method to balance costs and real-world validity, and present results clearly with concrete risk models and policy suggestions. He concludes that designing and disseminating evaluations with policymakers in mind is crucial for maximizing policy impact.

Kevin Wei
RAND

AI Supply Chains:

An Emerging Ecosystem of AI Dependencies

Sarah Cen examines how the AI industry's specialization has created complex interdependencies between actors. The ecosystem now spans model developers, hardware providers, cloud platforms, data managers, and downstream applications. Cen identifies four reasons these dependencies matter: choke points where single failures cascade through the system, market concentration where power accumulates, accountability diffusion as responsibilities spread across actors, and talent flows that reveal shortages and conflicts. Her Stanford team is building a graph-based tool to map these relationships using public data sources including SEC filings and press releases. The project deliberately relies on public data to surface what information remains hidden behind confidentiality clauses. Initial findings already show organizations disclose different information about the same relationships across different venues, highlighting gaps in transparency that could inform policy.

Sarah Cen
Stanford University

Hydropower:
The Missing Piece

Charles Yang argues that hydropower is an overlooked solution for meeting the escalating power demands of AI data centers. Yang evaluates hydropower against alternatives like nuclear, geothermal, and solar energy, emphasizing its key strengths: established technology, significant existing capacity, and potential for expansion. He points out that only 3% of US dams currently generate power, leaving room for growth through retrofitting and modernization. Yang analyzes hydropower's policy landscape, examining its connections to defense, regulatory, and energy policy. He closes with a historical parallel to World War II, when hydropower delivered industrial baseload power to shipyards, manufacturing operations, and aluminum production—proposing that AI presents an opportunity for hydropower's revival.

Charles Yang
Center for Industrial Strategy

Compute in America

Policy Playbook for Secure 5GW Clusters

Arnab Datta outlines the challenges of building five-gigawatt AI compute clusters in the US within five years. These facilities would match the output of several nuclear plants—infrastructure the US has struggled to build while Middle Eastern and Chinese competitors advance with state backing. The key barrier isn't energy availability but permitting delays caused by litigation risk, exemplified by the Cardinal Hickory transmission line's decade-long approval process. Datta proposes "special compute zones" where the government expedites energy permitting for companies that meet AI security standards. His policy toolkit includes faster DOE loan deployment, continued IRA tax credits, using the Defense Production Act to prioritize turbine orders for data centers, streamlined environmental reviews under NEPA, and converting retiring coal plants that already connect to the grid.

Arnab Datta
Institute for Progress

FAR.AI

Watch the full playlist from
Technical Innovations in AI Policy

Learn from AI safety experts in
academia, industry & governance.

hello@far.ai

company/far-ai

@FARAIResearch

@FARAIResearch

far.ai