

Pathways for Advancing AI Safety in Congress

Strategic Intelligence for the AI Governance Community

O. Beno, S. Bollinger, M.A. Celone, P. Courtney, M. Parker

Summary

This brief synthesizes a full-length research paper examining why AI safety arguments that command expert consensus fail to generate legislative traction in the US Congress. The core finding is that the problem is not one of persuasion but of institutional translation. Policymakers and AI safety experts each hold knowledge the other needs, but the infrastructure for exchanging it is underdeveloped. The paper maps the barriers to that exchange, documents what has and has not worked across US state legislatures and the EU, and identifies concrete pathways for building the institutional connective tissue between technical expertise and democratic governance.

THE CENTRAL INSIGHT

The research finds that the gap between expert consensus on AI risk and Congressional inaction is not primarily a problem of ignorance, capture, or irrationality on either side. AI safety experts understand risk trajectories that policymakers cannot yet see. Policymakers understand institutional constraints, constituent dynamics, and political feasibility that experts routinely underestimate. Both forms of knowledge are necessary for effective governance. Neither is sufficient alone.

The failure is in translation. Experts optimize for technical correctness: the response they believe the evidence warrants. Policymakers optimize for political feasibility: responses that are implementable, defensible, and aligned with constituent concerns. These are not competing values but complementary functions that a well-designed governance system would connect. The mechanisms for making that connection -- standards processes, advisory bodies, staff training, cross-sector coalitions -- exist in fragmented form but lack the coordination and sustained investment to function effectively.

The procedures documented in this research, from NIST frameworks to agency rulemakings to procurement conditions, are not workarounds for failed communication, they are the communication. They are how governance systems create space for expert knowledge to inform democratic decision-making without requiring either side to fully adopt the other's framework. Building this infrastructure is the strategic task.

WHY TRANSLATION FAILS

Most readers will recognize these dynamics from experience. The research's contribution is documenting how they interact as a reinforcing system.

Structural barriers fragment the receiving end: committee jurisdictions split AI across Commerce, Judiciary, Armed Services, and others. No independent technical capacity has existed since the Office of Technology Assessment's elimination in 1995. Short electoral cycles bias towards reactive governance. **Political economy dynamics** distort the information environment: industry functions as an extension of Congressional staff, and the asymmetry in resources between industry and the governance community constrains the range of policy options that reach legislators. **Cognitive divergences** mean the same information is processed differently: experts see compounding uncertainties as reasons for caution; policymakers see them as reasons to wait for clearer evidence.

WHAT THE EVIDENCE SHOWS ABOUT INDUSTRY STRATEGIES

Analysis of AI legislation in California, Colorado, Texas, Utah, and the EU AI Act reveals consistent patterns. Industry frames governance as a threat to innovation; claims technical infeasibility to block specific provisions (as with SB 1047's shutdown requirement in California); publicly advocates for harmonized standards while working to dilute unified frameworks in other jurisdictions (the same companies calling for US federal preemption lobbied to narrow the EU AI Act); and sustains engagement after legislation passes to delay

implementation or force revisions (Colorado's AI Act effective date was pushed back through post-passage lobbying). The most decisive outcome was Governor Newsom's veto of California's SB 1047 after intensive industry engagement.

Key Finding

Catastrophic risk framing consistently failed to outweigh immediate economic arguments in legislative contexts. In California, existential risk warnings did not shift the calculus against economic counter-arguments. Prescriptive mandates such as SB 1047's shutdown capability requirement provided concrete targets for opposition. The paper's cross-domain analysis of aviation, nuclear, and pandemic governance finds that performance-based frameworks, which define safety outcomes rather than prescribe technical implementation, have historically proved more tractable under similar institutional constraints.

WHERE TRANSLATION IS ALREADY WORKING

Despite the barriers, the research identifies openings where expert knowledge and policymaker priorities are converging.

Three areas of bipartisan convergence. National security evaluations (the Frontier Model Forum has called for testing programmes, mirroring bipartisan Congressional legislation), child safety (AI-generated CSAM is documented and existing laws provide enforcement hooks), and fraud prevention (AI-enabled election interference and deepfakes align with existing legal frameworks). These are present-day harms where expert knowledge and legislative appetite already overlap.

The gradual disempowerment bridge. Near-term harms and long-term catastrophic risks are not separate categories but different points on the same trajectory. The erosion of human agency, institutional capacity, and democratic oversight is observable now. When experts warn about catastrophic outcomes and policymakers focus on economic displacement or democratic integrity, they are describing the same dynamic at different stages. This is the most productive framing for connecting the two perspectives.

Administrative embedding. NIST frameworks, agency rulemakings, procurement conditions, and standards processes persist across electoral cycles and are less vulnerable to political shifts. Many are already partially active. These are the institutional channels through which expert knowledge can reach governance without requiring comprehensive legislation.

RECOMMENDATIONS

1. Lead with documented harms from current models. Terrorist recruitment, CSAM generation, and election interference are concrete, present-day, and align with existing legal frameworks. They represent areas where expert knowledge directly serves policymaker needs. Frame governance as a precondition for sustainable deployment and public trust. Propose outcome-based requirements rather than prescriptive mandates.

2. Build coalitions with sectors that need governance. Healthcare, finance, labor, and child protection organizations cannot scale AI adoption without governance frameworks. They need clear liability rules, auditing standards, and regulatory pathways as preconditions for deployment. These sectors bring established regulatory relationships, bipartisan credibility, and organized advocacy infrastructure. Align the community around a short list of targeted, winnable priorities these coalitions can support.

3. Invest in institutional translation mechanisms. The paper identifies specific channels: research fellowships placing governance expertise in policy institutions; long-term Congressional staff training programmes; voluntary pilot audits between willing labs and federal agencies; coordinated participation in NIST standards processes; and a distributed governance approach working through existing regulatory authorities rather than waiting for comprehensive legislation. These mechanisms are how expert knowledge becomes usable by democratic institutions.

4. Coordinate existing efforts. Fellowships, voluntary audits, state-level campaigns, NIST standards development, and international alignment are already underway in fragmented form. The bottleneck is coordination, not invention. The full paper provides detailed playbooks for eight specific policy objectives, each identifying relevant actors, procedural vehicles, and tactical steps.

5. Maintain momentum beyond voting cycles. Colorado's experience shows that industry engagement intensifies after legislation passes. Plan for sustained engagement beyond the vote, build relationships with executive offices and agency implementation teams, and budget for implementation beyond campaigning from the outset.

FROM ANALYSIS TO COORDINATION

The research is direct about what it asks of both sides. AI safety experts need to engage seriously with how policy is made: building advocacy capacity, learning legislative rhythms, and framing arguments in terms that connect to the priorities policymakers already hold. Policymakers and the institutions that support them need to become more absorptive of technical expertise: through staff training, advisory structures, and sustained engagement with the research community. Neither side can do this alone, and neither side currently does it well enough.

The full paper provides the detailed evidence base, including policy playbooks with specific actors, procedural vehicles, and tactical steps. This brief is the starting point. The paper is the working document and makes recommendations for forward-looking work.

Full paper: Beno, O., Bollinger, S., Celone, M.A., Courtney, P., & Parker, M. (2026). *Pathways for Advancing AI Safety in Congress*. Arcadia Impact AI Governance Taskforce.