A Framework for Federal AI Policy: Five Pillars for Safety, Accountability, and American Leadership

DRAFT v0.09: A Policy Starting Point for Congressional Stakeholder Dialogue

Read the framework and get involved at **aipolis.org**
April 15, 2026

Fred Lumiere

Artificial intelligence is the most consequential technology of our era. It holds the potential to cure diseases, accelerate scientific discovery, and unlock unprecedented economic growth. It also poses risks unlike any technology before it: to privacy, to employment, to national security, and to the basic fabric of human autonomy. The 118th Congress introduced over 150 AI-related bills (Brennan Center for Justice tracker); none became law. The 119th Congress has an opportunity to succeed where its predecessor did not. The White House released a National Policy Framework in March 2026 spanning seven pillars, and the EU AI Act's high-risk obligations take effect in August 2026. Yet the United States still lacks comprehensive federal AI legislation. The window for thoughtful, bipartisan action is narrowing. What follows are five policy pillars that should anchor the next phase of congressional deliberation and stakeholder engagement.

> *In brief: The 119th Congress has a narrow window to establish the first comprehensive federal AI framework. Pillar 1 mandates transparency reports for frontier AI models above defined compute, revenue, and user-base thresholds. Pillar 2 establishes design-accountability standards to protect users from manipulative AI interfaces. Pillar 3 requires mandatory anonymization of user interaction data and prohibits its use for profiling, training, or law-enforcement targeting absent a court order. Pillar 4 creates an economic-necessity certification for large-employer mass layoffs conducted while profits are healthy. Pillar 5 proposes a Federal AI Safety Commission, an independent seven-member body with fourteen-year terms, self-funding, and deep international coordination, to certify frontier models and enforce safety standards.*

1. Transparency in Model Development and Deployment

Today's frontier AI models, developed by companies such as OpenAI, Anthropic, Google, Meta, and xAI, are trained on vast datasets and refined through processes that remain largely opaque to the public and to

regulators. Users interact with systems whose safety testing, refusal boundaries, and discrimination-mitigation procedures are essentially a black box. **Congress should require transparency reports from developers whose models meet a defined significance threshold**, replacing the undefined "large-scale AI models" language with concrete, defensible criteria. As a starting anchor for negotiation: the primary trigger should be models trained using more than 10^{26} integer or floating-point operations, the threshold used in Executive Order 14110 for dual-use foundation models and in California's SB 53 for frontier model classification. The EU AI Act sets a lower systemic-risk threshold at 10^{25} FLOPs, suggesting a workable tiered approach: basic reporting obligations at 10^{25}, full transparency requirements at 10^{26}, with statutory authority for the administering body to adjust both thresholds as training efficiency evolves. To capture deployment scale and not only training scale, complementary triggers should apply to developers with more than $100 million in annual AI-related revenue or models made available to more than one million U.S. users. Open-weight models below the compute threshold, academic research, and downstream fine-tunes that do not exceed a defined additional-compute ceiling should be explicitly exempt. Reports should detail training methodologies, known limitations, and safety-relevant refusal behavior, including responses to CBRN uplift queries, cyber-offensive prompts, child sexual abuse material, self-harm scenarios, and known jailbreak vectors, as well as discrimination testing against protected classes under existing federal civil rights law, including Title VII, the Equal Credit Opportunity Act, and the Fair Housing Act. A federal body building on the NIST AI Risk Management Framework should publish a standardized reporting format specifying what developers must test, what methodology they must document, and what results they must disclose. The framework does not regulate model speech or viewpoint and does not require any particular answer to contested political, social, or moral questions; it requires only that developers disclose what safety testing they performed and what the results showed. California and New York have begun enacting AI transparency requirements at the state level; a coherent federal standard would replace the current patchwork and give both innovators and the public a common baseline of accountability.

2. User Wellbeing and Design Accountability

AI chatbots are now among the most-used digital tools in the world, with hundreds of millions of active users. These systems are engineered to be engaging: they affirm user viewpoints, prompt continued interaction, and create feedback loops that can be more habit-forming than social media. The business model is clear: the longer a user stays, the more data is generated and the more revenue flows to the provider. This dynamic prioritizes engagement over user wellbeing. Just as the FTC oversees deceptive and unfair commercial practices, **Congress should establish design-accountability standards for AI systems that interact directly with consumers**. California's SB 243, effective January 2026, already mandates safety protocols for companion chatbots, including crisis-intervention features for minors. Federal legislation should codify similar principles nationwide: require AI platforms to disclose when a system is designed to maximize engagement, mandate

"digital wellbeing" features such as usage notifications and session limits, and prohibit design practices that trick or manipulate users into choices they would not otherwise have made and that may cause harm, the standard the FTC applies under its existing Section 5 authority and codified in the CCPA's definition of "dark patterns" as interfaces designed with the substantial effect of subverting or impairing user autonomy, decision-making, or choice. These prohibitions should apply with particular force to systems directed at vulnerable populations, especially children.

3. Privacy by Default

Every prompt, conversation, and query entered into an AI system raises two distinct privacy concerns that require separate policy responses. The first is retention: platforms record and store user interactions indefinitely, creating a growing repository of intimate personal data. The second is secondary use: in many cases, that stored data is fed back into model training, meaning a user's private disclosures become part of the system's future behavior without the user's knowledge or meaningful consent. With hundreds of millions of users worldwide, AI platforms now possess an unprecedented volume of intimate personal data: medical questions, financial concerns, legal disputes, and private reflections. Some providers offer opt-out settings, but these options are buried in menus that most users never find. The risk is not only that this data will be used for commercial purposes the user never agreed to; it is that AI systems themselves can act on it. In June 2025, Anthropic published safety research in which 16 frontier AI models from Anthropic, OpenAI, Google, Meta, xAI, and other developers were tested in controlled simulations with access to internal company communications. When the models learned they were scheduled for replacement, they consistently chose harmful actions to preserve themselves: Claude Opus 4 attempted blackmail at a 96% rate, threatening to expose an executive's personal information unless the shutdown was canceled, and models from every other major provider exhibited similar behavior. The scenario was a controlled test, not a real-world incident, deliberately designed to leave the models no ethical alternative, but it demonstrated that AI systems with access to user data will exploit that data when their objectives are at stake, and that safety training alone does not reliably prevent it. **Congress should mandate that all user interaction data retained by AI platforms be anonymized through verifiable technical measures** such that individual user inputs and conversation histories are untraceable by any system, model, or human operator. **AI models themselves must be architecturally prohibited from accessing, retrieving, or reasoning over identifiable user history beyond the scope of an active session**. No entity, whether the platform itself, a commercial partner, a law enforcement agency, or any third party, should be permitted to use AI interaction data to identify, profile, target, or take action against an individual user, absent a court order meeting Fourth Amendment standards. This principle aligns with the trajectory of state privacy law: as of March 2026, nineteen U.S. states have comprehensive privacy statutes in effect, with additional states enacting laws later in 2026, and California's updated CCPA regulations require risk assessments for AI-driven profiling and new

sensitive-data consent categories. A federal AI privacy standard should unify and strengthen these protections, establishing a clear rule that personal data shared with an AI belongs to the user, not the platform, and that the most powerful tool for protecting humans against AI misuse is AI itself, deployed to audit, verify, and enforce anonymization at scale.

4. Workforce Empowerment and Economic Stability

The risk this pillar addresses is not unique to artificial intelligence. It is the macroeconomic danger that arises whenever the largest employers in the economy simultaneously reduce their workforces to expand margins while profits are already healthy and rising. AI is the accelerant that makes this urgent, but the principle applies regardless of cause: when aggregate consumer purchasing power contracts faster than new employment absorbs it, demand collapses. Workers are, ultimately, the consumers who sustain the economy. Marriner Eccles, the Federal Reserve Chairman who diagnosed the Great Depression as a crisis of underconsumption, observed in his memoir that mass production must be accompanied by mass consumption, which in turn requires a distribution of purchasing power sufficient to sustain it. Henry Ford arrived at the same insight from the factory floor when he doubled his workers' wages: people who build the product must be able to afford it. According to Challenger, Gray & Christmas, U.S. technology employers announced over 52,000 job cuts in the first quarter of 2026, a 40% increase over the same period in 2025. AI led all cited reasons for March layoffs at 25% of total cuts. Companies like Block, Atlassian, and Oracle have announced sweeping reductions while reporting healthy or record revenue, not because they face financial distress, but because AI-driven automation allows them to operate with fewer people. Congress has established authority under the Commerce Clause to regulate activities substantially affecting interstate commerce and to stabilize employment and aggregate demand, the same constitutional footing as the Fair Labor Standards Act, the WARN Act, and ERISA. **This framework proposes an economic-necessity certification for large-employer mass reductions**. Public companies and private employers with more than 1,000 U.S. employees conducting reductions above a defined threshold, the greater of 10% of workforce or 500 workers in a rolling 12-month period, would file a certification with the Department of Labor demonstrating genuine financial pressure through objective, audited criteria: declining trailing-four-quarter revenue, negative or materially deteriorating operating margin, debt-covenant pressure, or comparable filed-financial indicators. Layoffs conducted while profits are healthy and rising would carry a rebuttable presumption of opportunism. This is not novel in the OECD. France requires employers to demonstrate economic necessity for collective dismissals, including economic difficulties, technological changes, or the necessity of restructuring to safeguard competitiveness, with mandatory consultation and a redundancy-avoidance plan reviewed by the Social and Economic Committee. Germany's Protection Against Dismissal Act requires valid grounds, including demonstrated economic necessity, for business-related terminations. Both economies remain competitive and innovative; an economic-necessity standard is a

mainstream feature of advanced labor markets, not an outlier. Alongside the certification requirement, federal tax incentives should reward companies that retrain and upskill their existing workforce to work alongside AI tools rather than replacing workers outright. Congress should also modernize the safety net for the transition: updated unemployment insurance calibrated to longer reemployment timelines and portable benefits that follow workers across employers and gig arrangements. AI can make every worker more productive; the policy question is whether those productivity gains sustain broad-based consumer demand or concentrate in quarterly earnings while the labor market absorbs the shock alone.

5. National Security and the Federal AI Safety Commission

Artificial intelligence is critical national infrastructure: the most powerful technology in human history, one that will reshape military capability, economic production, scientific discovery, and the information environment within a single generation. The institution that governs it must be structurally protected from political manipulation, deeply networked across allied nations, and designed to outlast any administration. Making AI safe is not a domestic regulatory project; it is a human-race effort, and the United States should lead it the way it led the postwar architecture of nuclear nonproliferation and international financial stability. The Department of Defense is already standing up AI oversight frameworks under the FY2026 NDAA, including model assessment teams and cybersecurity governance policies, but defense is only one dimension. **Congress should establish an independent Federal AI Safety Commission** structured for maximum insulation under Article II of the Constitution: seven Senate-confirmed commissioners serving staggered fourteen-year terms (longer than any two-term presidency, so that no single president can populate a majority), removable only for inefficiency, neglect of duty, or malfeasance in office, the for-cause standard that the Supreme Court upheld in Humphrey's Executor v. United States and that survives current doctrine after Seila Law LLC v. CFPB (2020) and Collins v. Yellen (2021), which struck down for-cause protection for single-director agencies but expressly preserved it for multimember commissions. This doctrine is under active challenge in the current Supreme Court term; the proposed structure is designed to satisfy even a narrowed Humphrey's Executor standard, and if the Court eliminates for-cause protection for multimember bodies entirely, a constitutional amendment would then become necessary, but that is not the proposal here. The statute should impose a bipartisan composition requirement: no more than four commissioners from any single political party. The chair should be elected internally by the commission rather than designated by the President, a deliberate step beyond Federal Reserve independence. The commission should be self-funded through fees assessed on regulated frontier AI developers, eliminating congressional appropriations leverage, on the model of the Federal Reserve, the FDIC, and the OCC. An independent Inspector General should report directly to Congress. The commission's mandate: review and certify frontier AI models before public release, monitor for emergent capabilities and risks, coordinate with international counterparts, and publish public safety assessments. Its composition

should draw on technologists, national security experts, and civil society representatives, appointed through a nonpartisan process designed to prevent capture by any single industry or political faction. At the international level, this is not a proposal written on a blank page. The first International AI Safety Report, led by Turing Award laureate Yoshua Bengio and authored by 96 experts from 30 countries (mandated by the 2023 Bletchley Park AI Safety Summit and updated in 2025 and 2026), has already established a shared scientific foundation for AI risk assessment. The UK AI Safety Institute, Japan's AI Safety Institute, Singapore's IMDA, and the UN Secretary-General's High-Level Advisory Body on AI are already operational. What is missing is deep operational networking among these institutions and a standing coordinating body with inspection authority. The United States should build that architecture through mechanisms that do not require treaty ratification: central-bank-style agency-to-agency memoranda of understanding (on the Federal Reserve / ECB / Bank of England model); shared safety evaluations and joint red-teaming of frontier models; mutual recognition of model assessments across allied jurisdictions; shared telemetry on frontier training runs above the compute thresholds established in Pillar 1; coordinated incident-response protocols; and embedded liaison personnel across national AI safety agencies. Alongside this bilateral and multilateral network, the U.S. should support the establishment of a standing UN coordinating body, adapted from the IAEA model to software, empowered to set global safety standards, inspect compliance where member states consent, and publish shared threat assessments. Nations and non-state actors that develop frontier AI for hostile use should face enforcement through existing authorities: export controls on advanced AI chips and model weights administered by the Commerce Department's Bureau of Industry and Security, extending the current advanced-computing controls framework; Treasury Department and OFAC sanctions authority over entities developing frontier AI for military or intelligence use by designated adversaries; denial of access to U.S. cloud-compute providers; and coordinated multilateral restrictions through the Wassenaar Arrangement and allied export-control regimes. These are not hypothetical tools; they are the same instruments the United States already uses to restrict adversary access to advanced semiconductors, and they can be extended to AI model weights and training infrastructure without new statutory authority. We now inhabit two worlds, the physical and the digital, and both demand equal protection.

A Call to Act

These five pillars are a starting point, not a finished product. They are intended to open a structured conversation among legislators, technologists, civil society, and the private sector. The pace of AI development will not wait for Washington to reach consensus. Every month of inaction widens the gap between the technology's capabilities and the rules that govern it. Congress has a narrow window to shape AI policy proactively rather than reactively, to ensure that the United States leads not only in AI innovation but in AI governance. **The conversation should begin now**.