Governing Artificial Intelligence:
A Risk-Based Framework for a U.S. Agency for AI Deployment

Aarnav Verma, Amen Salha, Lex Dallas, Michael Burns, Milenka Men

Executive Summary

The rate of progress in artificial intelligence has outpaced the institutions meant to govern it, moving quickly from research labs into the heart of healthcare, finance, surveillance, and defense. The United States currently relies on a fragmented mix of sector-specific federal oversight and inconsistent state laws, operating without a unified system to determine when deployment is safe or legitimate. This proposal calls for the creation of a U.S. Agency for Artificial Intelligence, a dedicated federal body that would regulate the deployment, but not the development, of AI through a tiered, risk-based framework. Under this model, low-risk systems would face transparency requirements; higher-impact systems would require audits and human oversight; and the most high-impact systems would undergo pre-deployment review by an interdisciplinary board.

Background

Artificial intelligence has moved quickly from being a niche area of academic research to becoming one of the most influential technologies shaping the global economy, governance, and national security. For many years, AI development largely took place in universities and research labs where scholars focused on improving machine learning systems, natural language processing, and computer vision. Over the past decade, however, improvements in computing power, access to massive datasets, and advances in neural networks have pushed artificial intelligence out of the research environment and into everyday use. Today, AI systems are increasingly embedded in digital platforms and infrastructure, shaping how businesses operate, how governments make decisions, and how people interact with information.

One of the main reasons for this shift has been the rise of generative AI and foundation models. These systems can generate human-like text, images, and code while also performing complex analytical tasks. Foundation models are trained on extremely large datasets and can be adapted to many different uses, which has expanded the range of industries that artificial intelligence can impact. Technology companies such as OpenAI, Google, Microsoft, and Anthropic have invested billions of dollars into developing these models and integrating them into widely used products and software platforms (McKinsey Global Institute 2023). As these systems continue to improve, many economists and policymakers view artificial intelligence as a major driver of productivity and future economic growth.

At the same time, artificial intelligence has become an important area of international competition. Governments increasingly see leadership in AI as critical for economic strength, technological leadership, and national security. The United States currently maintains advantages in research institutions, venture capital investment, and semiconductor technology that support AI development. However, China and several other countries have launched national strategies aimed at strengthening their domestic AI industries and reducing dependence on foreign technology (Lee 2018). As a result, many analysts now describe the current moment as a global race for artificial intelligence leadership.

The importance of AI also extends beyond economic competition and into national security. Artificial intelligence technologies are expected to shape areas such as cybersecurity, intelligence analysis, military systems, and critical infrastructure. The United States National Security Commission on Artificial Intelligence concluded that leadership in AI will play a significant role in determining future economic and military power (National Security Commission on Artificial Intelligence 2021). Because of this, governments are increasing investments in AI research, infrastructure, and talent development.

At the same time, the rapid expansion of artificial intelligence has raised important questions about governance and regulation. AI systems now influence financial markets, labor markets, information systems, and public decision making. Without appropriate oversight, these technologies may introduce risks such as algorithmic bias, misinformation, and economic disruption (OECD 2019). Policymakers therefore face the challenge of encouraging innovation while also ensuring that AI systems are developed and used responsibly.

Artificial intelligence is no longer a niche technological frontier. Instead, it is becoming a core layer of infrastructure shaping markets, governance, and global competition.

AI Impacts

Artificial intelligence is beginning to shape many aspects of society, creating both significant opportunities and serious risks. As AI systems become more integrated into healthcare, research, business operations, and government decision-making, their influence continues to grow. These technologies have the potential to increase productivity, expand access to services, and accelerate scientific discovery (McKinsey Global Institute 2023; Stanford Institute for Human Centered Artificial Intelligence 2024). At the same time, the rapid deployment of AI systems has raised concerns about fairness, misinformation, security, and accountability. Understanding both the benefits and risks of artificial intelligence is therefore essential when considering how governments should approach regulation.

One of the most promising areas where AI is already making an impact is in medical diagnostics. Machine learning systems can analyze medical images and large health datasets in ways that allow patterns to be detected earlier than through traditional methods alone. AI models have shown strong results in identifying conditions such as cancer, cardiovascular disease, and diabetic retinopathy (Topol 2019). These tools are not intended to replace physicians, but rather to support clinical

decision-making by helping doctors identify potential health concerns earlier. As healthcare systems face increasing demand and limited resources, AI-assisted diagnostic tools have the potential to improve both efficiency and patient outcomes.

Artificial intelligence is also contributing to productivity and scientific research across many sectors of the economy. AI systems can analyze large datasets, summarize complex information, generate computer code, and assist with technical tasks that previously required significant human effort. Researchers are increasingly using AI to model complex systems and identify solutions to difficult problems. In areas such as drug discovery and materials science, AI has helped researchers identify promising compounds more quickly than traditional research methods alone (McKinsey Global Institute 2023). In addition, AI technologies are expanding accessibility by enabling speech-to-text tools, automated translation systems, and image recognition software that assists individuals with visual impairments (World Economic Forum 2023).

AI is also beginning to play a role in accelerating scientific discovery itself. Advanced computational models are now used to analyze climate data, simulate biological processes, and predict complex chemical structures. One well-known example is AlphaFold, an AI system that predicts protein structures with remarkable accuracy and has significantly accelerated biological research (Jumper et al. 2021). These developments demonstrate how artificial intelligence can complement human expertise and expand the pace of innovation.

Despite these benefits, the rapid expansion of AI systems also introduces several risks. One concern involves algorithmic discrimination in automated decision-making systems. Because many AI models are trained on historical datasets, they may reflect existing social inequalities. As a result, automated systems used in hiring, lending, or criminal justice risk assessments may unintentionally reproduce biased outcomes (O'Neil 2016). Without transparency and oversight, these systems could reinforce existing disparities rather than reduce them.

Another growing concern is the spread of deepfakes and AI-generated misinformation. Generative AI systems can now produce highly realistic images, videos, and written content that may be difficult to distinguish from authentic information. This raises challenges for journalism, democratic institutions, and public trust in digital media (Chesney and Citron 2019).

Concerns have also emerged regarding AI use in healthcare decision-making. As artificial intelligence becomes more integrated into medical systems, questions about accountability and liability are becoming increasingly important. In recent years, lawsuits have been filed against healthcare organizations and insurers that relied on AI algorithms to make treatment or coverage decisions. In one case, families alleged that an AI model used by a health insurer incorrectly denied necessary medical care to elderly patients by relying on automated predictions rather than physician judgment (CBS News 2023). In addition, reports have documented cases where AI-assisted medical technologies contributed to surgical errors or misidentified anatomy, resulting in patient injuries and legal action against device manufacturers (Reuters 2026). These cases highlight the growing challenge of determining responsibility when AI systems influence medical decisions, particularly when multiple actors such as hospitals, physicians, and technology developers are involved (Maliha 2021).

AI technologies also raise concerns in national security and surveillance contexts. Some experts warn that advances in artificial intelligence could enable the development of autonomous weapons systems that operate with limited human oversight (Future of Life Institute 2023). In addition, the growing use of AI for surveillance and predictive policing raises questions about privacy, civil liberties, and the potential for disproportionate monitoring of certain communities (Richardson, Schultz, and Crawford 2019).

Taken together, these benefits and risks highlight why a risk-based approach to AI governance is necessary. Artificial intelligence has the potential to generate enormous social and economic benefits, but its rapid deployment also introduces meaningful challenges that must be addressed through thoughtful policy and regulatory frameworks.

Current Regulatory Landscape

AI oversight is distributed across a range of existing federal agencies that regulate AI only within the specific sectors already under their jurisdiction, while state governments expand existing policy to encompass the new risks of AI without a nationwide framework keeping the industry in check.

State-Level Policy

Despite the lack of a nationwide act from Congress, multiple states have acted to mitigate the risks of AI development and deployment. In 2024, the Colorado Artificial Intelligence Act (CAIA) established one of the first state-level frameworks, addressing "high-risk" AI systems used in employment, housing, healthcare, and other important decisions in daily life. The legislation places responsibilities on both developers and deployers of AI systems to reduce the risk of algorithmic discrimination while still encouraging competition and innovation. However, the law faced strong opposition from technology companies, leading to amendments and delays before its planned implementation in June 2026.

California has enacted multiple laws affecting AI deployment, including expanded provisions within the California Consumer Privacy Act and by the California Privacy Protection Agency. Additional proposed legislation in California has been modeled on aspects of the European Union's AI Act, including bills to address training data transparency for generative AI systems and requirements for watermarking AI-generated content.

Illinois has also been particularly active in regulating emerging technologies. The state's Biometric Information Privacy Act, passed in 2008, requires written consent before companies can collect biometric identifiers and allows individuals to sue companies that violate these rules. Illinois later enacted the Artificial Intelligence Video Interview Act (AIVIA), which requires employers to notify applicants when AI is being used, how it is being used, obtain consent, and limit who can view the videos. There are also additional proposals that would expand transparency requirements and treat algorithmic bias as a civil rights violation.

Besides Colorado, California, and Illinois, an increasing num-

ber of states have implemented narrower policies to address specific AI-related risks. For example, Texas regulates biometric data collection through the Capture or Use of Biometric Identifier Act. It has also established an Artificial Intelligence Advisory Council responsible for overseeing responsible AI development. New York City has adopted Local Law 144, which requires bias audits for any automated employment decision tools used during the hiring process. Together, these state initiatives demonstrate the growing relevance of AI regulation on the policy agenda, but these overlapping compliance regimes for companies operating nationwide simply contribute to a fragmented regulatory environment.

Federal Policy

Several federal agencies play distinct roles in AI oversight. The National Institute of Standards and Technology (NIST) developed the AI Risk Management Framework in 2023, which aims to establish national technical standards and coordinate safe research activities across the AI ecosystem. Colorado has validated the use of the framework by incorporating it into its implementation of CAIA. However, NIST does not possess direct regulatory authority, and its framework is mostly advisory and relies on voluntary adoption by companies.

The Federal Trade Commission (FTC) functions as the primary federal enforcement body addressing AI-related harms in commercial markets. The FTC uses its authority to investigate deceptive or unfair business practices that involve generative AI systems. This includes issues such as fraud and misinformation affecting consumers. While the FTC has enforcement capability related to automated decision systems, its authority is limited to consumer protection and competition issues rather than comprehensive AI governance.

In healthcare, the Food and Drug Administration (FDA) regulates AI used in medical devices and clinical decision tools. AI-powered diagnostic systems and machine-learning medical software must undergo regulatory review similar to other forms of medical technology before being deployed in healthcare settings. The FDA has developed pathways for evaluating these machine-learning systems, but its jurisdiction remains limited to healthcare applications, consistent with agency-specific patchwork regulation.

Beyond these agencies, numerous federal departments and regulatory bodies address AI within their respective policy domains. The Department of Health and Human Services has provided guidance on how the Health Insurance Portability and Accountability Act (HIPAA) applies to AI systems processing protected health information. In addition, the Consumer Financial Protection Bureau has issued statements warning that AI systems used in lending or consumer finance must comply with anti-discrimination laws. These expanded interpretations rely on existing legal frameworks rather than new AI-specific legislation.

Likewise, multiple long-standing civil rights and consumer protection laws are interpreted in application to automated systems. The Civil Rights Act prohibits employment discrimination even when decisions are made by automated systems. The Fair Credit Reporting Act similarly regulates the use of discriminatory systems in credit scoring and background checks. The Equal Credit Opportunity Act prohibits discrimination in lending decisions, while the Fair Housing Act applies its principles to automated systems used in housing decisions. Together, these laws indirectly govern AI regulation through preexisting legal precedent.

Governmental Use of AI

Another defining feature of U.S. AI governance is the close relationship between government institutions and private technology companies. Unlike some countries that develop state-owned AI systems, the United States relies heavily on private firms to build and deploy advanced technologies. As a result, the government functions simultaneously as a regulator, customer, and strategic business partner. Companies are tasked with developing the systems, while government agencies are expected to fund research, set safety standards, and regulate technological applications downstream.

Several companies illustrate this public-private partnership model. Palantir develops data analytics and AI systems used by the military, law enforcement, and intelligence agencies, and it has several contracts with the Department of Defense, Immigration and Customs Enforcement, and the CIA. Similarly, AI developers such as Anthropic and OpenAI have become major stakeholders in policy discussions, especially due to matters of national security associated with advanced AI systems.

In 2025, the Pentagon's AI office awarded contracts worth up to $200 million to companies including OpenAI, Anthropic, Google, and xAI to develop AI systems for national security missions. These systems are intended to support functions such as battlefield operations and cybersecurity. However, collaboration between the government and AI developers is not straightforward. Anthropic reportedly imposed restrictions on how its AI systems could be used by the Pentagon, including prohibitions on mass surveillance of citizens and restrictions on use in fully autonomous weapons systems. These limitations ultimately contributed to tensions between the company and government agencies. As a result, some agencies shifted toward other companies, including OpenAI, for similar capabilities.

Overall, the current regulatory landscape for AI in the U.S. is decentralized and highly sector-specific. Federal agencies regulate AI only within their traditional jurisdictions, state governments are pushed to experiment with new legislative frameworks, and the government appears as a simultaneous consumer and partner with private technology companies. While this system allows for flexibility and innovation, it also creates significant coordination challenges compared to other countries with a more centralized approach to AI governance.

Recent Developments

Recent developments with regard to the use of artificial intelligence systems in national security have made this proposal more pertinent than ever before. Reports that Anthropic's Claude models were used in the Maduro Venezuela raid suggest that frontier model providers are already becoming intertwined with federal operational activity. In late February and early March 2026, the DoW's conflict with Anthropic became a focal point in the governance debate after Anthropic resisted demands for

broader military use of its systems, particularly where such use could implicate mass surveillance or autonomous weapons. Anthropic has framed its position through its Responsible Scaling Policy. OpenAI has since entered into an agreement with the DoW. OpenAI published a statement describing its agreement with the Department of War and said the arrangement incorporates restrictions tied to existing defense guidance, including verification and testing requirements for autonomous and semi-autonomous systems. It has since been updated to include a provision explicitly prohibiting domestic surveillance, autonomous weapons use, and autonomous decision-making.

As frontier AI systems become embedded in intelligence and defense workflows, private model governance mechanisms such as acceptable-use policies, contractual restrictions, and safety frameworks begin to shape the scope of government capabilities. The relationship between AI developers and state institutions therefore raises questions about the appropriate balance between corporate governance and public oversight. At the federal level, these firm-level disputes are unfolding alongside a broader shift in U.S. AI policy toward competitiveness, centralized national strategy, and resistance to a fragmented state-by-state regulatory regime. White House materials from late 2025 and early 2026 emphasize economic leadership, national security, and the development of a national AI action framework, while also criticizing certain state-level AI laws as obstructive.

Policy Proposal

This proposal recommends the creation of the U.S. Agency for Artificial Intelligence (USAAI). The agency would be a federal department dedicated specifically to overseeing the governance and deployment of artificial intelligence systems. Similar to the Food and Drug Administration's regulatory approach to pharmaceuticals and medical technologies, the agency would operate through a tiered risk-based regulatory framework that categorizes AI systems according to the level of risk they pose to public safety, civil rights, and national security. The central function of the agency would be certification for deployment, rather than regulating the research and development of AI models themselves. In this framework, companies, universities, and laboratories would remain free to develop AI technologies without government approval, addressing concerns that regulation could slow innovation or weaken U.S. competitiveness in the global AI race. However, before an AI system can be released publicly or deployed in real-world applications, it would need to receive certification from the USAAI depending on its assigned risk tier. Structurally, the department would be led by a Secretary of AI, or similarly broad title, who would serve as the primary liaison between the agency, Congress, and the executive branch. At the same time, the agency would function similarly to institutions such as the Federal Reserve, being independent within the government rather than independent of it. That is, it would operate under congressional mandates while maintaining insulation from short-term political pressures when making technical regulatory decisions.

A key component of this framework would be the establishment of a Pre-Deployment Review Board responsible for evaluating all AI systems prior to certification. Other structural changes may be necessary to accommodate the growing rate of AI models seeking deployment. Nonetheless, this board would consist of an interdisciplinary panel of experts drawn from fields including computer science, engineering, medicine, safety engineering, ethics, philosophy, law, civil rights, and national security. The board would ensure that AI systems are evaluated not only for technical functionality but also for their broader societal, ethical, and security implications. The board would also have final decisions on the integration of artificial intelligence within government institutions. This review structure ensures that powerful AI systems receive careful scrutiny before entering public use, while still preserving the freedom to innovate and develop new technologies within research environments.

Members of the Pre-Deployment Review Board would be nominated by the President and confirmed by the Senate and House to ensure democratic accountability. To reduce political influence and promote long-term stability, board members would serve for terms of between two and four years. Appointment criteria would require interdisciplinary expertise. Conflict-of-interest safeguards would limit appointments from individuals with significant financial ties to companies developing AI systems under review. However, a persistent concern remains that there are no regulatory systems that can fully bypass the influence of lobbying, making it difficult to design an institution that is entirely independent from political and corporate pressure.

Tiered System

We categorized AI usage into the three tiers based on the following questions: What kind of harm could happen with its use? Is the harm mostly inconvenience, or could it affect someone's job, freedom, health, or safety? How serious would that harm be? Can a human realistically catch and correct mistakes? Does the system touch protected rights or coercive state power?

Tier 1.

What belongs here. AI systems used for convenience, creativity, or low-stakes assistance. Oftentimes, using AI within these fields results in mistakes that are unlikely to seriously harm a person's rights, opportunities, health, or safety. Their outputs may be wrong, but they usually do not create irreversible harm. Users also tend to be able to notice mistakes and disregard them. The idea is to focus on accountability and licensing, but not overburden these companies or government officials while overseeing AI usage.

Examples. Consumer chatbots, writing and productivity assistants, customer service helpers, basic recommendation engines, and scheduling or note-taking tools, if not used for medical purposes.

Policy. The primary focus is on light obligations, mainly transparency requirements. Requirements include clear disclosure that the user is interacting with AI; a basic explanation of what the system does; notice of known limitations; consumer safety disclosures; acknowledgement that direct information from chats will not be sold; a mechanism for reporting harmful or deceptive outputs; recordkeeping by the company; and authority by the department to suspend, recall, or prohibit deployment.

Tier 2

What belongs here. AI systems used to help make decisions that materially affect people's access to opportunity or treatment, but where humans are still expected to retain oversight. While these systems do not usually create immediate physical danger, they can systematically affect a person's life. They can deny jobs, loans, housing, educational opportunities, and public benefits. The harm is more serious because it can be hard to detect, and it may fall unevenly across varying racial, gender, and religious groups.

Examples. Hiring and resume-screening systems, educational grading or admissions support tools, credit scoring, housing screening tools, insurance pricing or claims triage tools, and government benefit eligibility tools.

Policy. While Tier 1 focuses primarily on transparency, Tier 2 would focus on pre-deployment procedural accountability. Requirements include all policies from Tier 1 in addition to the following: bias testing; documentation of training data sources and limitations; an annual independent audit and third-party assessment by the department to retain certification; a human oversight requirement for final decisions; and a right to explanation or appeal for affected individuals.

Tier 3: High Risk AI Systems

What belongs here. AI systems deployed in sectors where a failure could seriously threaten life, bodily safety, or constitutional and civil rights. This tier also includes AI systems where the state may use AI in coercive ways. These systems can produce catastrophic or irreversible harm such as wrongful arrests, loss of life, or discriminatory surveillance. The FDA's regulatory model for AI-enabled medical devices is useful here because it treats safety and effectiveness as core conditions for market usage.

Examples. Autonomous weapons systems, predictive policing tools, facial recognition for law enforcement or mass surveillance, medical diagnostic or treatment systems, and border enforcement and national security systems.

Policy. Models in this tier would be subject to the strongest controls, focusing on continuous review and oversight of users. Requirements include all policies from Tiers 1 and 2 in addition to the following: safety and ethics protocols; product review by an independent, interdisciplinary Pre-Deployment Review Board; and oversight by the department of all users.

Enforcement would include holding search engines accountable if they allow uncertified AI systems to be found and used on their platforms.

Sources

References

[1] Bommasani, Rishi, et al. *On the Opportunities and Risks of Foundation Models*. Stanford Center for Research on Foundation Models, 2021. https://arxiv.org/abs/2108.07258

[2] CBS News. "Families Sue Health Insurer Over AI-Driven Care Denials." 2023. https://www.cbsnews.com/news/unitedhealthcare-ai-lawsuit-care-denials/

[3] Chesney, Robert, and Danielle Citron. "Deepfakes and the New Disinformation War." *Foreign Affairs*, 2019.

[4] Congressional Research Service. *Artificial Intelligence and National Security*. 2023. https://crsreports.congress.gov

[5] Future of Life Institute. "Autonomous Weapons and Artificial Intelligence." 2023. https://futureoflife.org/open-letter/autonomous-weapons-ai/

[6] Glacis. *Guide to State AI Laws*. https://www.glacis.io/guide-state-ai-laws

[7] Jumper, John, et al. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature*, 2021.

[8] Lee, Kai-Fu. *AI Superpowers*. 2018.

[9] Maliha, A. "Artificial Intelligence and Medical Liability." *Journal of Law and Medicine*, 2021. https://pmc.ncbi.nlm.nih.gov/articles/PMC8452365/

[10] McKinsey Global Institute. *The Economic Potential of Generative AI*. 2023. https://www.mckinsey.com

[11] National Security Commission on Artificial Intelligence. *Final Report*. 2021. https://www.nscai.gov

[12] O'Neil, Cathy. *Weapons of Math Destruction*. 2016.

[13] OECD. *Recommendation of the Council on Artificial Intelligence*. 2019.

[14] OpenAI. *GPT-4 Technical Report*. 2023. https://arxiv.org/abs/2303.08774

[15] Richardson, Rashida, Jason Schultz, and Kate Crawford. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data." NYU Law, 2019. https://ainowinstitute.org/publication/dirty-data-bad-predictions

[16] Reuters. "Reports of Surgical Errors Linked to AI-Assisted Medical Devices." 2026. https://www.reuters.com/investigations/ai-enters-operating-room-reports-arise-botched-surgeries-misidentified-body/

[17] Stanford Institute for Human Centered Artificial Intelligence. *AI Index Report 2024*. https://aiindex.stanford.edu/report/

[18] Topol, Eric. *Deep Medicine*. 2019.

[19] U.S. Artificial Intelligence Innovation Ecosystem. 2023. https://www.ai.gov

[20] University of Denver. "Colorado Pumping the Brakes on First-of-Its-Kind AI Regulation to Find a Practical Path Forward." https://www.du.edu/news/colorado-pumping-brakes-first-its-kind-ai-regulation-find-practical-path-forward

[21] White House. *America's AI Action Plan*. 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[22] World Economic Forum. *AI for Accessibility*. 2023.

[23] *Anthropic Responsible Scaling Policy v3*. https://www.anthropic.com/news/responsible-scaling-policy-v3

[24] *Fortune*: "OpenAI Pentagon Deal, Anthropic Designated Supply Chain Risk, Unprecedented Action Could Damage Its Growth." 2026. https://fortune.com/2026/02/28/openai-pentagon-deal-anthropic-designated-supply-chain-risk-unprecedented-action-damage-its-growth/