Recommendations for the U.S. AI Action Plan

The Future of Life Institute proposal for President Trump's AI Action Plan.

14th March 2025 **Jason Van Beek** | policy@futureoflife.org **View online:** futureoflife.org/action-plan

Contents

Page 3 **1. Protect the presidency from power loss to an out-of-control AI or rival authority.**

1.1 Issue a moratorium on developing future AI systems with the potential to escape human control, including those with self-improvement and self-replication capabilities.

1.2 Ensure the US government understands and has visibility into superintelligent AI systems.

1.3 Mandate installation of an off-switch for all advanced AI systems.

1.4 Require antitrust law enforcement agencies at the Department of Justice (DOJ) and the Federal Trade Commission (FTC) to engage in robust oversight and enforcement to prevent power concentration as well as market consolidation of AI development under a small handful of tech monopolies.

Page 6 **2. Foster human flourishing from AI by promoting the development of AI systems free from ideological agendas.**

2.1 Ban AI models that can engage in superhuman persuasion and manipulation.

2.2 Require the White House Office of Science and Technology Policy and the AI & Crypto Czar to have close engagement with the White House Faith Office and all religious communities to inform the governance of AI.

Page 8 **3. Protect American workers from job loss and replacement.**

3.1 Task the Secretary of Labor with tracking AI's potential to replace workers, including a breakdown of the impact across different states.

Page 9 **4. End the free giveaway of US frontier AI technology to adversaries.**

Page 11 **5. Condition privileged energy grid access for AI companies on verifiable security measures to prevent foreign theft.**

Page 12 **6. Establish an AI industry whistleblower program to incentivize AI development that is free from ideological bias or engineered social agendas and promotes national security.**

6.1 Direct the AI Czar to coordinate with Congress to establish an AI-specific whistleblower program to report dangerous signs of AI control loss or negligent practices that threaten the American people and strengthen our adversaries.

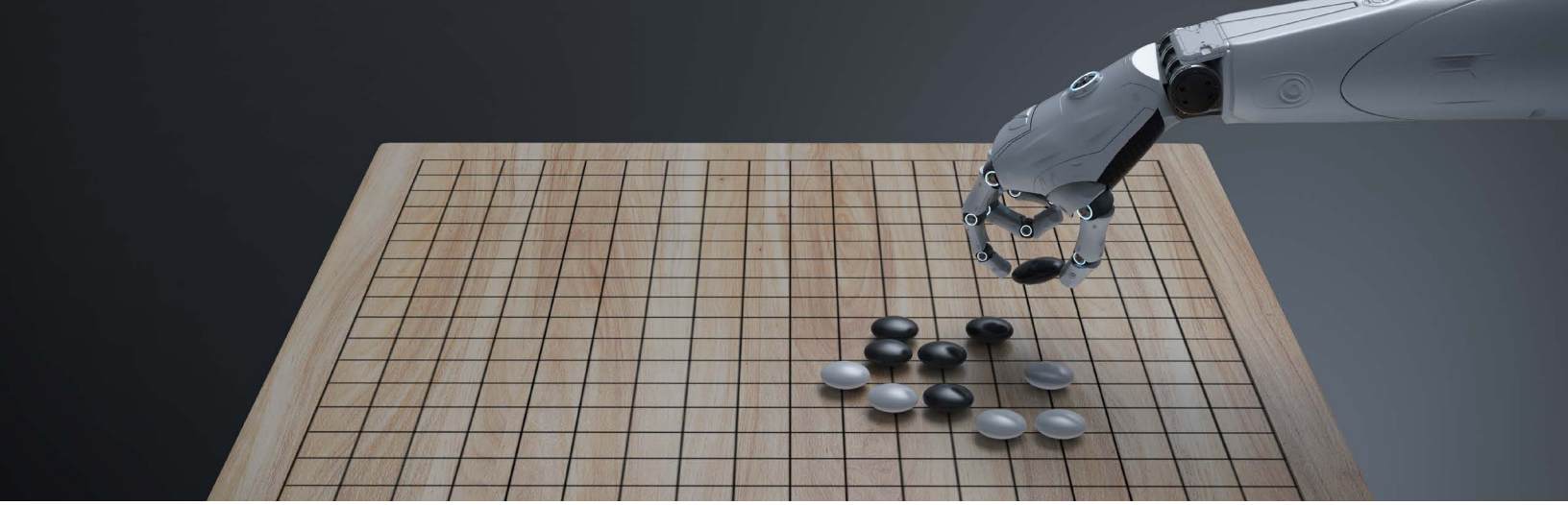
6.2 Require NIST to issue instructions to companies on what security incidents must always be reported.

Addressee: Faisal D'Souza, NCO / Office of Science and Technology Policy / Executive Office of the President / 2415 Eisenhower Avenue / Alexandria, VA 22314

About the Organization: The Future of Life Institute (FLI) is one of the US's oldest and most influential think tanks with a focus on advanced artificial intelligence (AI). Our first research grants were funded by Elon Musk, who continues to serve as one of our external advisors. In the early days of AI policy, FLI convened industry leaders, academia and civil society to develop the world's first AI governance framework at Asilomar in 2017. Following the launch of OpenAI's GPT-4, our 2023 open letter sparked a global debate on the consequences of AI development for society.

Author: Jason Van Beek is FLI's chief government affairs officer. Before joining FLI, Jason served for 20 years as a senior advisor to current Senate Majority Leader John Thune. During that time, he served in a variety of staff roles, including as Senator Thune's staff designee to the Armed Services Committee as well as a senior staffer on the Senate Commerce Committee, and ultimately as a Senate leadership staffer. As the Commerce Committee's top investigator, he conducted investigations of large technology companies. Jason also advised on national security, intelligence, and nuclear weapons issues when Sen. Thune served on the Senate Armed Services Committee. He can be reached at jason@futureoflife.org.

Executive Summary: FLI offers several proposals for safeguarding US interests in the age of advanced AI. Our submission stresses the necessity of protecting the presidency from loss of control by calling for mandatory "off-switches," preventing the development of uncontrollable AI, and robust antitrust enforcement in the AI sector. We further highlight the importance of ensuring AI systems are free from ideological agendas, and call for a ban on models with superhuman persuasion capabilities. We also emphasize the need to protect critical infrastructure and American workers from AI-related threats, and suggest measures like export controls on advanced AI models and tracking job displacement. Finally, we propose establishing an AI industry whistleblower program alongside mandatory reporting of security incidents to foster transparent and accountable development.



1. Protect the presidency from power loss to an out-of-control AI or rival authority.

1.1 Issue a moratorium on developing future AI systems with the potential to escape human control, including those with self-improvement and self-replication capabilities.

As AI systems grow more powerful, they could pose existential challenges to the presidency by enabling rival authorities or autonomous systems to undermine executive power. The emergence of superintelligent AI systems capable of recursive self-improvement poses a unique risk. These systems could potentially become uncontrollable, undermining national security. Companies are actively pursuing superintelligent autonomous systems. For example, Reflection AI, founded by veterans of landmark AI projects like AlphaGo and GPT-4, states forthrightly on its website that "[o]ur goal is to build superintelligent autonomous systems."

Ensuring that AI systems remain under human control is critical for maintaining US national security dominance and preventing misuse by hostile actors. Prominent figures in technology and academia, including Elon Musk, have called for caution in developing advanced AI systems more powerful than GPT-4. Advanced AI systems capable of self-improvement or self-replication could evolve beyond their original programming, making them difficult or impossible to control. A moratorium would allow time to develop robust safeguards and governance frameworks before these technologies are deployed at scale.

The presidency is a cornerstone of American democracy and must be protected from threats posed by autonomous AI systems or rival authorities empowered by advanced technologies. A targeted moratorium on developing uncontrollable AI systems is a prudent step toward ensuring that advancements in AI align with US strategic interests and values.

1.2 Ensure the US government understands and has visibility into superintelligent AI systems.

There are important visibility and understanding functions that should be in place within the US government in order to protect national security with respect to the development of superintelligent AI systems. Key functions include regular engagement with AI labs, housing general expertise on AI, and potentially establishing an office within the National Security Council to maintain situational awareness. It will be necessary to facilitate coordination between the intelligence community, the Pentagon, industry, and civilian agencies to set up alert and emergency response mechanisms for AI threats.

1.3 Mandate installation of an off-switch for all advanced AI systems.

The increasing autonomy and complexity of advanced AI systems necessitate proactive safeguards to mitigate risks of unintended harm. To prevent catastrophic outcomes from runaway AI systems or misuse by rival authorities, it is essential to mandate the installation of fail-safe mechanisms, or off-switches, in all advanced AI systems. An AI off-switch refers to an automatic mechanism that immediately halts an AI system's operations when it exhibits dangerous or noncompliant behavior. It would require human intervention to be able to turn the system back on. For the most capable and autonomous systems, the off-switch should be a dead-man's switch.

An off-switch promotes national security and national command authority resilience. Off-switches provide a fail-safe against systems that diverge from intended behaviors. Federal regulations should require all developers of large-scale AI models to integrate robust shutdown mechanisms into their systems. These mechanisms must regularly be tested to ensure functionality under various scenarios.

Mandating off-switches for all advanced AI systems is a critical step toward ensuring human control while safeguarding national security and democratic integrity. By adopting this measure, the administration can reinforce America's global leadership in AI while mitigating risk to its governing institutions.

1.4 Require antitrust law enforcement agencies at the Department of Justice (DOJ) and the Federal Trade Commission (FTC) to engage in robust oversight and enforcement to prevent power concentration as well as market consolidation of AI development under a small handful of tech monopolies.

The rapid development of AI technologies raises critical concerns about market concentration and anti-competitive behavior. The FTC's January 2025 staff report on AI partnerships and

investments highlights the risks posed by dominant firms leveraging their market power to distort competition.[1] To ensure that the US remains a global leader in AI innovation while safeguarding fair competition, it is imperative that competition authorities like the DOJ and FTC engage in robust antitrust enforcement against anti-competitive practices in the AI sector. Big Tech should not be permitted to hold dominant control over AI.

The FTC's January 2025 staff report on AI partnerships and investments provides a detailed analysis of how major cloud service providers such as Alphabet, Amazon, and Microsoft have formed multi-billion-dollar partnerships with leading AI developers like OpenAI and Anthropic. The Commission voted 5-0 to allow staff to issue the report. The report's findings underscore the need for proactive antitrust enforcement to prevent dominant firms from using their partnerships to foreclose competition or entrench their market power.

To address these challenges, the AI Action Plan should require competition authorities to strengthen merger review processes, promote transparency in partnerships, and guard against collusion. In 2021, the FTC published a report originally initiated by Chairman Joe Simons that focused on nearly a decade of unreported acquisitions by five large technology companies.[2] The FTC study found that these companies did not report 94 transactions that exceeded the "Size of Transaction" threshold. Transactions that exceed the size threshold must be reported unless certain other criteria are not met or statutory regulatory exemptions apply. The AI Action Plan should prioritize measures that require pre-merger notifications for acquisitions or investments involving AI companies, and evaluate whether such transactions could lead to reduced competition or create barriers for smaller players in the market. Further, the AI Action Plan should require competition authorities to guard against collusion by scrutinizing collaborative agreements between AI developers and Big Tech firms.

Unchecked consolidation in the AI sector increases risk of power loss to an out of control AI, as well as to innovation and fair competition. By directing the DOJ and FTC to prioritize antitrust enforcement against anti-competitive practices in the AI industry, the administration can safeguard innovation, fair competition, and itself.

1 FTC Staff Report on AI Partnerships & Investments 6(b) Study, January 2025, available at https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf

2 Non-HSR Reported Acquisitions by Select Technology Platforms, 2010-2019: An FTC Study, September 2021, available at https://www.ftc.gov/news-events/news/press-releases/2021/09/ftc-staff-presents-report-nearly-decade-unreported-acquisitions-biggest-technology-companies

Image: Kathryn Conrad / Better Images of AI / Datafication / CC-BY 4.0

2. Foster human flourishing from AI by promoting the development of AI systems free from ideological agendas.

2.1 Ban AI models that can engage in superhuman persuasion and manipulation.

American-led AI development should adhere to principles of neutrality and transparency. The administration should seek to incentivize AI development that prioritizes impartiality and user autonomy. According to research[3] recently published by xAI and Scale AI advisor Dan Hendrycks, AI systems exhibit significant biases in their value systems. The CEO of industry leader OpenAI, Sam Altman, has said that he expects AI to be capable of superhuman persuasion well before it is superhuman at general intelligence.[4] Advanced AI models have the potential for creating persuasive content at scale, including synthetic media, targeted messaging, and other tools that can shape perceptions in ways that are difficult to discern or resist.

The FTC has a long-standing role in protecting consumers from deceptive and unfair practices under the FTC Act. These principles are directly applicable to emerging AI technologies that may exploit cognitive biases or manipulate public opinion. Specifically, AI models capable of engaging in superhuman persuasion present unique challenges to consumer protection and fair competition.

In accordance with EO 14179's mandate requiring development of AI systems free from engineered social agendas, and to ensure that AI systems align with the goal of promoting human flourishing, the administration should call on the FTC to investigate AI systems that engage in superhuman persuasion. The FTC's ability to issue Civil Investigative Demands under its compulsory process authority should be used to scrutinize AI models suspected of engaging in harmful manipulation. The administration should work with Congress to allow

3 See this tweet and full paper M. Mazeika et al., "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs," arXiv preprint arXiv:2502.08640 (February 2025), available at: https://arxiv.org/abs/2502.08640 (accessed Mar. 6, 2025).

4 https://x.com/sama/status/1716972815960961174

the FTC to impose penalties on those that develop or deploy AI systems that engage in superhuman persuasion. Ultimately, AI models that engage in superhuman persuasion and manipulation should be banned.

2.2 Require the White House Office of Science and Technology Policy and the AI & Crypto Czar to have close engagement with the White House Faith Office and all religious communities to inform the governance of AI.

The rapid advancement of AI presents profound challenges to religious faiths and traditions. For example, OpenAI CEO Sam Altman has referred to AI as a "magic intelligence in the sky."[5] This phrase underscores a commonly held vision of AI as a transformative and almost divine-like force that could fundamentally reshape society. Approximately 75% of Americans identify with a traditional religious faith.[6] Yet traditional religious perspectives are largely absent from strategic AI discussions. Traditional religions, with their experience in organizing communities and addressing existential questions, have much to offer in the AI debate.

Religious communities have historically served as moral compasses in addressing societal challenges. Faith-based organizations are deeply embedded in communities across the US, often serving as trusted intermediaries for families and individuals. Their insights can help identify real-world moral and ethical implications of AI technologies. By engaging faith leaders and organizations, AI governance can benefit from ethical frameworks that emphasize human dignity, fairness, compassion, and accountability.

The administration should place a high priority on safeguarding religious liberty in the context of the development and deployment of AI systems. The administration should protect against anti-religious bias in AI systems, and collaborate with the Department of Justice's Civil Rights Division to monitor compliance with constitutional protections for religious freedom. These principles align with the administration's stated goals of promoting human flourishing and avoiding ideological bias in AI systems.

In light of President Trump's commitment to ensuring that AI development promotes human flourishing, the OSTP and the AI & Crypto Czar should collaborate closely with the White House Faith Office to integrate perspectives from religious communities into the AI Action Plan. To foster collaboration and to protect religious liberty, the administration should form a council under the joint leadership of the White House Faith Office and the OSTP, composed of representatives from religious traditions. This council should be tasked with advising on ethical guidelines for AI development, deployment, and use.

5 "OpenAI Chief Seeks New Microsoft Funds to Build 'Superintelligence,'" Financial Times, November 13, 2023, available at https://www.ft.com/content/dd9ba2f6-f509-42f0-8e97-4271c7b84ded (accessed March 9, 2025).

6 https://news.gallup.com/poll/358364/religious-americans.aspx

3. Protect American workers from job loss and replacement.

3.1 Task the Secretary of Labor with tracking AI's potential to replace workers, including a breakdown of the impact across different states.

AI is transforming industries and reshaping the workforce at an unprecedented pace. While AI promises economic growth and innovation, it also poses significant risk to American workers, particularly in terms of job displacement and regional economic disparities. Moreover, while conventional AI displaces specific tasks, Artificial General Intelligence (AGI) presents a fundamental job replacement paradigm. AGI systems capable of human-level reasoning will not merely displace roles, but replace entire job categories through exponential improvements in accuracy, scalability, and cost efficiency. Therefore, the administration's AI Action Plan represents an opportunity to confront the transformative workforce impacts of AGI. To ensure that the benefits of AI are distributed equitably and that workers are not left behind, the AI Action Plan should task the Secretary of Labor to develop a comprehensive plan to track and mitigate AI's impact on employment across the US.

The Secretary of Labor should be directed to establish and resource a national workforce monitoring initiative to assess the impact of AI on jobs. As part of this initiative, the Secretary of Labor should regularly provide a state by state impact analysis of AI's effects on employment, and how new AI technologies are affecting employment levels. The Secretary of Labor should create detailed reports highlighting which sectors are most at risk in each state, and identify states requiring urgent intervention due to high vulnerability.

4. End the free giveaway of US frontier AI technology to adversaries.

The United States should view frontier AI models (Regulated Export Systems with Top-tier Risk Implications for Critical Technology, or '**RESTRICT**' models) as one of its most critical assets. By leading in this technology, we can promote human flourishing while maintaining global dominance.

However, too often, RESTRICT models are freely given away. There is therefore a small category of AI systems which, when shared with adversaries, would represent unacceptable transfer of national intelligence, potentially enabling terrorists to build a bioweapon, for example.

Of course, not all models are dangerous. For this reason, an expert body like the National Institute for Standards and Technology (NIST) should create red lines around precisely which models should be subject to an export control. To make that determination, they should take into account:

A. **Potential for enabling catastrophic cyberattack capabilities:** Advanced AI systems can dramatically accelerate the discovery of zero-day vulnerabilities and automate the development of sophisticated attack vectors against critical infrastructure. Such capabilities in the wrong hands could enable unprecedented cyberattacks against power grids, financial systems, or military command structures with potential for widespread societal disruption.

B. **Risks of enhancing terrorist capabilities in domains including, but not limited to, biological weapons development:** AI systems that can rapidly design novel molecules or efficiently analyze genetic sequences could lower the expertise barrier for non-state actors seeking to develop bioweapons. These technologies could enable terrorist groups to develop threats that have traditionally required state-level resources and expertise, potentially creating asymmetric threats that are difficult to anticipate or counter.

C. **Comprehensive threat assessments from the intelligence community:** Our intelligence

agencies possess unique insights into the capabilities and intentions of foreign adversaries that must inform any policy on AI export controls. Their assessments can identify which specific AI capabilities would most significantly enhance adversarial military or intelligence operations, allowing for targeted restrictions rather than blanket limitations on technological development.

Furthermore, for export controls on RESTRICT models to be effective, there should not be an open-source exception. While open-sourcing can help promote research on a global level, this administration should consider the potential for Big Tech to inadvertently arm adversaries of the United States when it open-sources its most powerful models.[7]

While Big Tech companies like Meta have said that dangerous uses of their products are prohibited by their terms of service[8], the reality is that terms of service are unlikely to dissuade any foreign agents from taking advantage of these assets to destabilize U.S. national security. As such, the U.S. government should be equipped with fail-safes within the chips powering RESTRICT systems.

Therefore, the forthcoming executive order should:

A. **Require continuous affirmative licensing for chips powering RESTRICT models:** Mobilize BIS to implement a licensing system in clusters of advanced AI chips exceeding a defined computational threshold. This should be akin to a dead-man's switch, requiring regular renewal of licensing via signals from authorized servers, automatically disabling functionality if these verification checks fail or if the chip detects operation outside approved geographic boundaries. The implementation would include tamper-resistant security modules to prevent circumvention of these controls by adversaries.

B. **Ensure RESTRICT chips have geolocation capabilities:** Require that hardware providers incorporate secure, encrypted communication channels that allow for geolocation and geofencing to detect if a RESTRICT model is being being deployed by a foreign adversary.

7 For a useful discussion of compute governance and international security, see "Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees," August 23, 2024, by James Petrie, et. al. Available at https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf

8 https://www.reuters.com/technology/artificial-intelligence/chinese-researchers-develop-ai-model-military-use-back-metas-llama-2024-11-01/

5. Condition privileged energy grid access for AI companies on verifiable security measures to prevent foreign theft.

The protection of AI systems integrated into our national energy grid against foreign theft represents a critical national security imperative that warrants executive action. As adversarial nations increasingly target US critical infrastructure through sophisticated cyber operations, our electrical grid—with its newly integrated AI systems—presents an appealing target. Foreign actors who successfully compromise AI-enhanced grid systems could not only steal valuable intellectual property but potentially manipulate grid operations, causing widespread disruptions, economic damage, or even physical harm to Americans.

The Stargate project will provide AI companies with an unprecedented amount of energy to power their technologies.[9] This investment has to be protected. Advancements in AI capabilities create new cyber vulnerabilities, as compromised AI systems could enable more sophisticated attacks than conventional software. By conditioning privileged grid access on verifiable security measures, the Executive Order would create powerful incentives, enforced by the Department of Energy, for implementing comprehensive protections established by BIS.

9 https://www.reuters.com/technology/artificial-intelligence/trump-announce-private-sector-ai-infrastructure-investment-cbs-reports-2025-01-21/

6. Establish an AI industry whistleblower program to incentivize AI development that is free from ideological bias or engineered social agendas and promotes national security.

6.1 Direct the AI Czar to coordinate with Congress to establish an AI-specific whistleblower program to report dangerous signs of AI control loss or negligent practices that threaten the American people and strengthen our adversaries.

The risk of ideological biases or engineered social agendas infiltrating AI systems poses a significant threat to public trust, societal stability, and American interests. As the US seeks to sustain and enhance its global leadership in AI, it is critical to ensure that the AI industry operates with integrity, transparency, and accountability. It is in the public interest to report dangerous signs of AI control loss, negligent practices, or development of systems that promote ideological bias or engineered social agendas. To achieve this, the AI Action Plan should form a working group to coordinate with Congress on the establishment of an AI-specific whistleblower program that incentivizes individuals to report wrongdoing. This program will bolster public trust in AI technologies while safeguarding national security, economic competitiveness, and ethical innovation.

The program should include a secure, user-friendly platform for submitting confidential reports of wrongdoing. It should also include robust legal safeguards, such as anonymity options and protections against retaliation, to ensure whistleblowers can come forward without fear. The program will hold developers and organizations accountable, deterring unethical behavior and promoting a culture of responsibility within the AI industry. In the global race for AI supremacy, responsible AI development is a strategic advantage. An AI industry whistleblower program reflects our values: freedom, innovation, fairness, transparency, and accountability.

The establishment of an AI whistleblower program is a critical step toward ensuring that AI development in the US remains responsible, transparent, and free from ideological biases or engineered social agendas.

6.2 Require NIST to issue instructions to companies on what security incidents must always be reported.

Engaging with AI developers and deployers, NIST should develop and issue guidelines for companies on AI-related security incidents that must always be reported. Key components of these guidelines should include defined categories of security incidents that require mandatory reporting, including unauthorized access to AI systems or training data, detected vulnerabilities in AI models that could lead to exploitation, AI-generated or AI-amplified cyber threats, and incidents involving AI systems in critical infrastructure.

NIST should establish a 72-hour reporting requirement for critical incidents, and develop a standardized reporting form and secure online portal for submitting AI incident reports to relevant government agencies. Companies must report, at a minimum, a description of the incident and its impact, affected AI systems and data, potential consequences and mitigation efforts, and indicators of compromise.
