# CHAPTER 4

## Trump 2.0 and the Global Role of the United States in AI Safety Governance

**Tony Onazi Oche & Christopher Isike**
Department of Political Sciences, University of Pretoria, South Africa

### Introduction

The United States (US) is the global leader in advanced AI development and design. This dominance is due to the high concentration of frontier AI companies and data centres within its territory. Some argue that for the US to sustain this dominance, it is critically important that these AI companies adopt an ethical development approach that incorporates safety protocols (Ramjee and Selinger 2025: 122). Based on recent government policies, one can argue that policymakers in Washington recognise the necessity for AI safety. Over the past five years, through the establishment of executive orders, standards and laws, the US has developed a relatively rule-based AI ecosystem. However, the government also has to consider the national interests of the country as far as AI development is concerned. In addition to these competing interests, safety policy gaps sometimes occur due to the existence of an AI black box (a complex system that 'mysteriously' generates its own intelligence) (Hendrycks 2024:117). Consequently, policymakers grapple with the challenge of regulating a rapidly evolving industry, often lacking clarity regarding the object of regulation. Despite these factors, the pace of AI safety policymaking still significantly depends on governments' commitment to AI safety.

The importance of AI safety requires that AI safety governance is assessed separately from a government's general AI governance agenda. A close analysis of the Biden and Trump administrations exemplifies the significance of executive power in national and global AI safety. Both administrations, although with slightly different viewpoints, have issued Executive Orders (EOs) to advance

American national interests in AI and address safety concerns. However, it is not clear which administration has focused more on AI safety relative to current projected risks (Hine 2024:49). The assessment of a country's performance in AI safety is often problematic due to conceptual issues. This is often as a result of analysts and scholars using the broader 'AI governance' framework to analyse AI safety governance. However, AI governance covers both AI development and AI safety. In contrast, this chapter uses an AI safety governance lens to analyse the policies of the Biden and Trump administrations with the goal of measuring the trajectory of AI safety governance in the US between 2011 and 2025.

The analysis focuses on policies that are directly linked with the administrations; as such, it excludes parliamentary laws and bills. Thus, it focuses on EOs, and the standards and regulations that draw legitimacy from EOs. Two reference points for the analysis are, firstly, the decision of President Donald Trump to discontinue the Biden administration's Executive Order 14110 (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) and, secondly, the recent adoption of an AI Action Plan. These executive 'soft laws' aim to foster an AI ecosystem that promotes responsible and safe AI deployment, while also focusing on promoting US dominance in AI development.

This chapter examines how federal AI policies differ between the Trump (both terms) and Biden administrations, while highlighting the significance of AI safety policymaking in the US for global AI safety. It bases its conceptual framework on AI safety governance, drawing from available policies found in databases, industry publications and scholarly materials to examine whether there has been a shift in AI safety governance over the years and, if so, what the implications are for global AI safety governance. The first part of the chapter explains AI safety governance as an analytical tool, while the second part addresses AI safety governance in the US. The third part examines global AI safety. Thereafter AI safety governance under the Trump and Biden administrations is examined., while the final section discusses the global implications of the US AI safety policies.

## AI safety governance: A conceptual clarification

The broader field of AI governance has gained traction over the last five years due to the rapid pace of AI development (Roberts et al 2024: 1275). AI governance frameworks aim to support AI development while addressing ethical and safety

questions, much like how the state manages its economic and strategic interests. The safety component of the frameworks focuses on the safe and responsible development and use of advanced AI models. It also attempts to balance national interests with safety regulations when developing AI policies at the local, regional and global levels.

A cursory look at scholarly definitions of AI terms shows that AI governance covers both development and ethical concerns. Reuel et al. (2024: 3) define AI governance as 'the processes and structures through which decisions related to AI are made, implemented, and enforced'. They further identify the need to differentiate Technical AI Governance (TAIG) from AI governance, since the former means 'technical analysis and tools for supporting the effective governance of AI'. This differentiation illustrates the challenges policymakers encounter while formulating AI policies. According to the authors, TAIG creates a methodology that describes the symbiotic relationship between machine learning engineers and policymakers in developing effective policies that are technically and socially relevant. While we share the opinion of Reuel et al. (2024) that policymaking needs to be complemented with technical knowledge in the AI policymaking domain, we further argue that this gap stems from conflating AI governance with AI safety governance.

While TAIG helps policymakers identify priority areas and governable aspects of technical AI, AI safety governance refers to the safety policies that emerge from the application of TAIG methodologies. This conceptualisation helps scholars grasp AI governance priorities. It also defines a unique theme for policymakers: devoting the necessary attention to AI safety issues, while avoiding the temptation to divert attention to other related but slightly different AI governance causes.

Addressing AI safety governance separately ensures a nuanced understanding of AI safety and reduces the tendency to lump safety issues under the broader AI governance framework. As discussed previously, using a general description can reduce focus on AI safety concerns. Thus, this book chapter proposes the adoption of 'AI safety governance' to describe frameworks and controls that are specifically designed to address AI safety concerns under the broader AI governance framework. This conceptualisation helps assess a government's policy commitment to safe and responsible AI more clearly. It also provides a clear pathway to identify and address AI safety policies and gaps, since AI

governance is a broader concept that encompasses all government and corporate policies aimed at encouraging the research, development and deployment of AI (Batool, Zowghi and Bano 2025:2).

The 'AI safety governance' framing supports the argument of Reuel et al. (2024: 24) for a closer relationship between technical and non-technical policymakers. The authors argue that most policymakers lack the technical knowledge to carry out what they refer to as "technical AI governance." To bridge this gap, it is necessary to study AI safety governance first as a distinctive field under AI governance, in order to identify the technical requirements for efficient policies that respond to AI risks. A multidisciplinary understanding of legal, ethical and AI technical knowledge is necessary to design policies that encourage rapid development yet also evaluate risks. However, policymakers fail to recognise the technical requirements involved in developing safety policies for the AI ecosystem. Therefore, in reality, those responsible for AI policymaking often lack a technical understanding of AI, while those with such expertise tend to concentrate on developing actual solutions and products instead of becoming involved in governance.

The AI safety governance framework provides clarity for policymakers by illuminating the unique skillset required to formulate policies that are specific and robust in tackling AI risks. It also provides analysts with the tools to evaluate a government's commitment to AI safety easily.

'AI safety governance' is a concatenation of 'AI safety' and 'governance' to form a standalone concept. It governs how 'practices and principles (AI safety)' are designed to protect vulnerable AI models from misuse and users from risk (Rudner and Toner 2021:1-2; McGrath, Jonker and Clark 2024). The technical requirement of assessing AI safety is challenging for policymakers from a non-technical background. In addition, AI safety standard solutions are adopted from other engineering fields, requiring a high degree of control of systems and their environment (Habli et al. 2025: 2). Buhl et al. (2025: 8) argue that the development of 'safety cases', a framework that explains how developers foresee and mitigate vulnerabilities, can make AI risks clearer to policymakers. In Figure 4.1 below, we illustrate the scope differentials between AI safety governance and AI Governance.

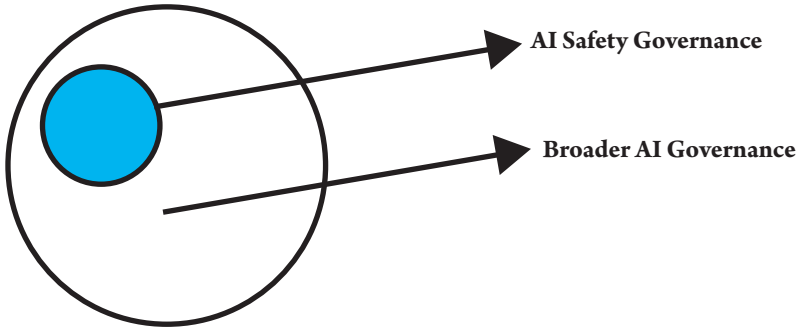



**Figure 4.1: Differentials between AI safety governance and AI governance. Source: Developed by the authors.**

Artificial Intelligence safety is generally the technical aspect of AI governance, but policymakers lacking technical understanding of AI systems may misappreciate the technical requirements of AI safety, which focuses on using computer science expertise to align AI systems with human values. Thus, conceptualising AI safety governance helps policymakers to appreciate its technical and distinctive characteristics compared to broader AI governance.

In this chapter, we use AI safety governance as a tool to identify and analyse AI safety policies in the US. As argued earlier, the AI safety policies drafted in Washington have a global effect due to US dominance in AI development. In the next session, we discuss the responsibilities and challenges of the US as a global leader in AI development, with reference to the AI safety governance concept.

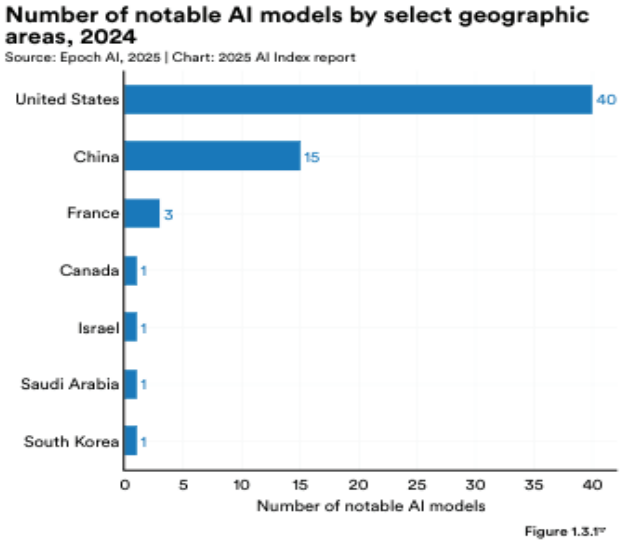
## The United States as a global leader in AI development: Responsibilities and Challenges

Several frontier AI companies design and produce their hardware as well as software in the US (*The Stanford University Artificial Intelligence Index Report* 2025; see Figure 4.2). These systems, especially the Large Language Model (LLM) variants, are developed using deep learning techniques. By means of these techniques, the LLMs are trained on a swath of internet data and afterwards fine-tuned through advanced machine learning techniques to provide in-context outputs in response to prompts during deployment (O'Brien, Ee and Williams 2023:12). Large language models, due to their capability to make sense of large swaths of data, could serve as general-purpose technology in the

future, meaning that they can be embedded in other systems to serve a range of purposes (Gruetzemacher and Whittlestone 2022: 3). The projected versatility of advanced AI increases its economic significance. This also makes advanced AI susceptible to risks that require effective governance throughout its entire life cycle.

The emergence of popular advanced AI models has increased the global deployment of LLMs for tasks that require considerable intelligence. Models such as OpenAI's ChatGPT, Google's DeepMind and Meta's Llama provide a range of functions—speech recognition, language generation and mathematical calculations—that help consumers to carry out research in diverse fields. However, because of these advantages, LLMs have also been integrated into military operations to increase efficiency and speed (Szabadföldi 2021: 161; Iqbal et al. 2023: 343). These gains, according to expert projections, could exacerbate human suffering. Some contend that the risks that AI pose could be catastrophic (Hendrycks, Mazeika and Woodside 2023: 6) or even existential.

These issues have inspired global advocacy for AI safety, and, given the position of leadership that the US occupies as evidenced by the compute capacity in Figure 4.2, the AI safety policies drafted in Washington could significantly shape the global landscape of AI safety governance. Furthermore, advanced models such as OpenAI's Operator (AI Agents) are currently only available in the US. Thus, due to their proximity to advanced AI companies and the technical knowledge necessary for developing safety policies for these models, policymakers in Washington have a unique obligation to regulate these advanced AI models before their release to other parts of the world, (Reuel et al. 2024: 1). This does not mitigate the importance of AI safety governance in other regions, but rather acknowledges the importance of creating AI safety policies during the development stage of LLMs.



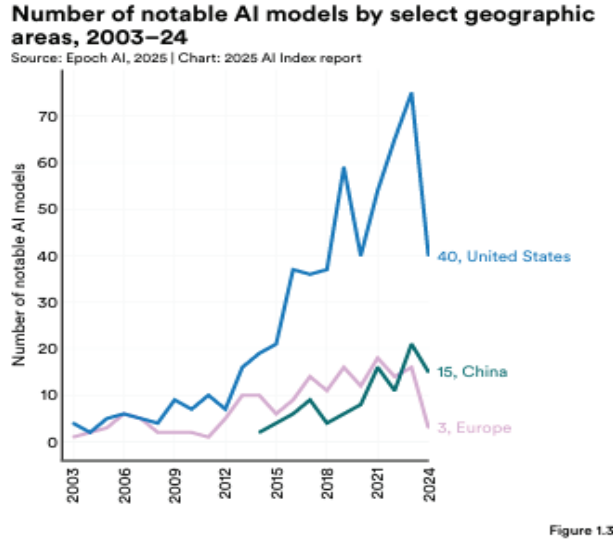




**Figure 4.2: Countries with the highest compute capabilities.**
**Source: *Stanford University Artificial Intelligence Index Report* (2025: 46).**

Therefore, the country that hosts a significant portion of frontier AI companies has the foremost duty to regulate these technologies in the development phase. Through regulation of private companies, the implementation of safety standards can be ensured (Heim et al. 2024: 2). Large language models are usually designed to learn during the pretraining stage. Since they are pretrained on internet data, they absorb both desirable and poor-quality data, which are difficult to remove from the memory of AI models (Maini et al. 2025: 2). Thus, pretraining data affects model performance and safety (Longpre 2024: 3245). During the fine-tuning stage, LLMs are trained on more structured and specific data sets that align with the values of the developer; however, as mentioned earlier, information absorbed during the pretraining stage remains in their memories (Hendrycks 2024: 168). This means that malicious users can manipulate models to provide outputs that they were not originally programmed to share with any end user (Shen et al. 2024: 1).

Frontier AI companies, aware of the risks posed by advanced AI models, have adopted safety policies to mitigate such risks. For example, these companies, have installed guardrails in the models to prevent unintended use; however, the guardrails have proven to be vulnerable (Hendrycks 2024: 147). They also often have an alignment and governance department that ensures the adoption of safety protocols during developmental stages. Similarly, government agencies such as the DoD recognise the importance of safety and have promised to adopt ethical practices when integrating AI into military operations (*US Department of State* 2023). Companies such as Palantir Gotham, OpenAI, Google and Meta have all invested in AI safety research (Delaney, Guest and Williams 2024: 2). While these commitments are still significantly inadequate, they indicate a consensus among stakeholders on the significance of safe and responsible AI models. This, therefore, shows the importance of safety policies for the US and other countries with a significant number of users. US dominance over the advanced AI global market means that it occupies a key leadership position in institutionalising AI safety standards for AI development.

## Global concerns over AI risks

Since the emergence of LLMs, there has been an increasing awareness of AI risks. The concerns are primarily about ethics and safety. Although traditional machine

learning models might have their associated risks, the AI models discussed here are frontier models, meaning the most advanced models at the time of writing (Salvador 2024: 5). Most concerns about advanced AI stem from its complex nature. While AI engineers combine 'randomness and human tweaking' to produce complex systems, they are unable to explain the internal operations of LLMs. The complexity of advanced AI has led to the emergence of AI alignment research, a field dedicated to elucidating and interpreting how AI systems achieve their intended goals without necessarily adhering to programmers' intended rules (Sabiens 2024). Advanced AI systems are developed by using a deep learning technique, which is the use of a computer program to build systems rather than actually programming systems. Advanced AI systems train neural networks, which are brain-like mathematical models, using a large amount of internet data during the pretraining stage. After that stage, the AI model goes through a training stage, where the model is fine-tuned using structured pieces of data to make predictions (outcome). Cortra (2021) provides an example of the challenge of advanced AI training and learning. According to the author, AI models are not intuitive in that, for example, they are unable to understand nuance in data sets the way humans do. As a result, they might pursue goals that do not align with the developer's values.

Advanced AI models, due to their opaque nature, can potentially pose certain risks. One of the most practical forms of risk could be unethical use by malicious actors (Marchal et al. 2024: 4). To limit the such risks, some experts propose implementing a responsible scaling policy (Hubinger 2025). However, the race to develop advanced models often prioritises corporate interests over safety. Even at the national level, there has been an increasing appetite recently to integrate LLMs into military operations to protect national interests. In the United States, Palantir Gotham (Egbert, Galis and Gundhus 2024: 2333) and Bolt & Bolt-M (Hambling 2024) are at the forefront of AI development for military use. OpenAI and Google have also pledged support for national security (Hambling 2024; Tucker 2025). Other risks could emerge due to competition between AI-leading countries (such as the United States and China).

**Table 4.1: AI incidents. Source: AI Incident Database Website (2025).**

| Incident | Date of Occurrence | Country | AI Model |
|---|---|---|---|
| Open AI's Agent called Operator spends $31 on eggs even though it was simply instructed by its human principal to check out the cheapest eggs on the internet. | 07 February 2025 | United States | Operator (ChatGPT 4) |
| Alleged misuse of AI-powered facial recognition technology leading to wrongful arrest | Published in the Washington Post on 13 January 2025 | United States | Several AI-powered facial recognition technologies |
| Brazil's Social Security AI Tool allegedly rejects social claims of Josélia de Brito, a former sugarcane worker. | Reported in | Brazil | Brazil's Meu Inss, a platform built by DataPrev leveraging AI to determine social benefits. |

To reduce risks that are already manifest, as evidenced in Table 4.1, international organisations, countries and companies have formulated AI safety policy frameworks. For example, the United States has developed standards, based on EOs, to ensure that AI companies adopt safety practices from the development stages of advanced AI, as we will see in the following sessions. Whether these policies are yielding results is contestable, although the overall aim is to reduce risks. Some AI safety policies aim to curb unethical use, while others concentrate on mitigating the risk of AI systems magnifying societal bias.

AI alignment is today a technical field that seeks to address AI risks in the development and corporate community. In practice, techniques such as fine-tuning, reinforcement learning with human feedback and unlearning are targeted at aligning AI models with the values of the developer (Qi et al. 2024: 8). Machine learning engineers refine a pretrained foundational model after training it on internet data. The overall aim of fine-tuning a model using a specific data set is to

customise it to respond to specific tasks. While this technique protects advanced fine-tunable AI models from responding to malicious prompts, they could nonetheless be susceptible to jailbreaks. Qi et al. (2024: 5) argue that both open-source models (ChatGPT 3.5 Turbo) and closed-source models (ChatGPT 3.5 Turbo) are vulnerable to malicious fine-tuning. Fine-tuning focuses on achieving specific tasks, while unlearning is a technique that seeks to remove harmful data from the memory of neural networks (Łucki et al. 2024: 1).

Despite the challenges of ensuring safe and responsible AI models, policymakers in the US have developed controls, risk assessments and prevention frameworks to guide AI companies and developers on pathways that mitigate AI risks. In this chapter, the focus is on the first and second Trump administrations and the Biden administration. The aim is to analyse the contrast between the two administrations in terms of their commitment to AI safety.

## AI safety governance in the Trump and Biden administrations

President Trump formulated AI governance policies during his first tenure, some of which addressed AI safety. In 2019, President Trump signed 'Executive Order 13859: Maintaining American Leadership and a Global Advantage in AI'. The main goal of the EO was to position the US as a global leader in AI to yield scientific discovery, economic growth and protection of national security. In addition, the EO stipulates that stakeholders should uphold technical standards while reducing barriers to safe testing and deployment of AI models, as well as increasing public trust in AI technologies (United States 2019). The National AI Initiative Act of 2021 reiterated the principles in Executive Order 13859. The Act allocates over $6 billion to finance various aspects of Executive Order 13859, among other matters. In 2020, President Trump signed 'Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government', which stipulates that federal agencies shall ensure a safe, secure, reliable and efficient use of AI (United States 2020).

During the Biden administration, many more AI safety policies seem to have been created. This could be due to several factors. The most prominent was the increasing awareness of AI risk and the advocacy from ethicists for a policy response. Additionally, the Biden administration's vision may have been aimed at integrating an ideology of fairness and equality into AI development

and use. One of the most popular AI safety policies of the Biden administration was the Blueprint for an AI Bill of Rights, which was released in October 2022. The framework, through its five principles, provides guidelines for AI companies and developers aimed at reducing risk during the development and deployment stages.

In October 2023, President Biden signed 'Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of AI' into law (United States 2023). The goal of the executive order was to ensure the safe and responsible development and use of AI models to maximise economic gains and minimise risks. Section 4 of the Act stipulates that the Secretary of Commerce, through the Director of the National Institute of Standards and Technology (NIST), should establish guidelines and best practices to ensure the safe development and deployment of advanced models. This framework includes the following: 'a. creating a resource to go along with the AI Risk Management Framework, NIST AI 100–1, specifically for generative AI; b. making a resource to work with the Secure Software Development Framework to add secure development practices for generative AI and dual-use foundation models; and c. starting a project to provide guidance and standards for assessing and checking AI abilities, especially in areas where AI could be harmful, like cybersecurity and biosecurity.'

While the Act indicates an awareness of the importance of safety, it creates an exception for national security in its provision for red-teaming testing processes for dual-use models. The NIST, in collaboration with the National AI Safety Institute, developed other notable frameworks during the Biden administration, as shown in the following table:

**Table 4.2: NIST Frameworks. Source: NIST Website (2025).**

| Framework/ Resources | Date of Issue | Focus Point |
|---|---|---|
| Four Principles of Explainable AI (Phillips, Hahn, Fontana, Yates, Greene and Przybocki 2021:3) | September 2021 | This article states the four principles of explainable AI to include: ‘<br>1. Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.<br>2. Meaningful: A system provides explanations that are understandable to the intended consumer(s).<br>3. Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.<br>4. Knowledge Limits: A system only operates under conditions for which it was designed,and when it reaches sufficient confidence in its output.’ |
| Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (Schwartz et al 2022). | 16 March 2022 | This article addresses bias as a critical AI threat and develops guidance for identifying and managing bias in three categories: systemic, statistical and human. In addition, it identifies the challenges of low-quality datasets, evaluation and human agency. The article also identifies ways of managing bias in advanced AI systems. |

| Framework/ Resources | Date of Issue | Focus Point |
|---|---|---|
| NIST AI Risk Management Framework (Tabassi 2023). | 26 January 2023 | The Framework is the foundational and leading NIST policy blue book on AI governance. It provides a detailed risk management guide adopting most definitions and practice direction from OECD recommendation on AI:2019; ISO/IEC 22989:2022. It lists characteristics of a trustworthy AI to include 'valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed'. |
| Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (Chandra, Dunietz Roberts 2024 | April 2024 | The Framework identifies image, video, voice and text as opportunities as well as threats to stability due to the possibilities of deepfakes and unethical use. It relies on existing practices and frameworks to mitigate the generation of illicit content. |
| Autio et al. (2024) | 24 July 2024 | This document derives its powers from Executive Order 14110. It provides insights into risks that are associated with advanced generative AI. It was developed as a supporting document to the AI Risk Management Framework. |
| Booth et al. (2024) | 26 July 2024 | This framework adopts Secure Software Development Practices Version 1.1, but narrows down its focus to data, development and integration of models in other software. It does not address cybersecurity risk generally at the deployment stage. As a result, the framework is designed to assist AI model producers, AI system producers and AI system acquirers in mitigating cybersecurity risks at the development stage. |

| Framework/ Resources | Date of Issue | Focus Point |
|---|---|---|
| US AI Safety Institute (2025) | January 2025 | This framework aims to provide guidelines for developers of foundational AI models to anticipate, mitigate and manage misuse, as well as comply with cybersecurity best practices to reduce model theft. |
| Vassilev et al. (2025) | March 2025 | This article identifies the difference between generative AI and predictive AI and develops a taxonomy and terminology for attacks that are accustomed to these types of adversarial machine learning. |
| NIST AI Risk Management Framework Playbook (National Institute of Standards and Technology n.d) | 20 March 2023 | The playbook is a voluntary document created to provide a guide for companies and organisations on how to achieve the outcomes set out in the AI Risk Management Framework. It focuses on how to govern, map, measure and manage risk. |
| NIST AI Risk Management Framework (AI RMF 1.0): An Explainer Video | 26 January 2023 | This video is a 6-minute 31-second video explanation of the NIST AI Risk Management Framework. It focused on methodologies and the content for several stakeholders. |

From the above table (table 4.2), it is clear that the Biden administration, in comparison to the Trump administration, adopted significantly more AI safety standards. Some argue that having so many policies can stifle innovation by increasing compliance costs, while having a reasonable number of standards encourage sustainable innovation (Tartaro, Smith and Shaw 2023: 4-7). Others argue that voluntary regulation, coupled with the AI development race, can motivate companies to comply with regulations as a competitive advantage (Han et al. 2022: 2). However, there is no clear evidence to show that safety standards in the Biden administration reduced AI development or innovation. The Trump administration also held the belief that Executive Order 14110 would stifle advanced AI innovation and development in the US, so it revoked the order

in January 2025 through another EO titled 'Removing barriers to American leadership in artificial intelligence'. While this shift will have a short-term effect, Kramer, Leavens and Scarlat (2024: 16) argue that executive orders are often at risk of being revoked by a successive administration. They also argue that one of the weaknesses of the Biden administration's policy was the delegation of implementation to NIST. This weakness can appear to be a strength, as delegation can engender sustainable compliance. In the final analysis, the revocation of Executive Order 14110 is more of a suspension than a revocation. A new administration can either revive it or develop new policies that address AI safety at a high level. Besides, the standards developed from the EOs adopted during the Biden administration are now part of US AI safety culture.

In addition to the 'Removing barriers to American leadership in artificial intelligence' EO, the Trump administration proposed a bill to place a moratorium on AI regulation through the '2025 congressional proposal for a moratorium on State AI regulation.' However, the US Senate voted against the bill. The Trump administration also released the AI Action Plan (*White House* 2025), which includes provisions aimed at promoting AI safety. The Action Plan spells out the government's plans to support initiatives that foster 'AI Interpretability, Control, and Robustness', risk evaluation mechanisms, mitigation and adoption of deepfake standards, promotion of a risk evaluation framework at the global level and adoption of safety standards and protocols to ensure biosecurity (*White House* 2025: 9—11). These policies show the support of the Trump administration for AI development, regardless of the cost of safety; however, a few of these policies, such as the AI Action Plan, for example, simultaneously address both development and safety interests in a single document. Indeed, from the approach of the Trump administration to broader AI governance, one can argue that the administration appears to significantly support development over safety. However, in reality, the EOs of the Biden administration that appear to promote AI safety were implemented through standards and controls under NIST frameworks. As a result, they have become institutionalised, making it difficult to change the safety culture through revocation of EOs during the Trump administration. However, reduced support for AI safety culture can weaken institutional frameworks and cause a decline in the culture of compliance in frontier AI companies.

## The importance of US AI safety governance for global leadership

The leadership position of the United States in global AI development makes its governance of AI safety crucial. This is clear from its commitment to the AI Action Plan, which addresses certain international AI concerns. As the leading country in AI development, the US intrinsically bears the responsibility to develop safety standards for AI frontier companies. The development stage is a strategic phase to implement most safety frameworks (Hilton et al. 2025: 2). Moreover, countries producing advanced AI systems often lead the way in deploying them and identifying their weaknesses. Thus, policymakers in the US develop standards and frameworks as part of their AI leadership responsibility. As we saw in the previous section of this chapter, most NIST frameworks target the development stages of advanced AI models. In contrast, countries that deploy rather than produce advanced models lack the capacity and jurisdiction to predict and regulate advanced models.

As a result, AI safety governance in the United States affects the global standard of responsible and safe AI models. Therefore, a decline in AI safety in the United States affects AI safety on a global scale. The United States is faced with a challenging situation in that it must decide between sacrificing safety to promote innovation and maintaining its status as the global AI superpower (Roberts et al. 2024: 1276). In addition, it is often challenging to reach a global consensus on the governance of new technologies, especially since a significant number of advanced models are also built in China. Ultimately, AI safety in the United States will shape, if not entirely determine, global AI safety governance; therefore, the AI safety paradigm in Washington is of global relevance.

However, despite the importance of the United States in AI safety governance, the risk posed by advanced AI models is not restricted to the US, even though most advanced models are built there. As exemplified in Table 1, AI risks can manifest in any part of the world, regardless of a country's capacity to produce advanced AI.

## Conclusion

The United States remains central to global AI safety due to its advancement in AI development. Thus, this chapter argued that AI safety policies drafted in

Washington have global relevance, as regulating the development of advanced AI determines the safety and responsible use of advanced models. In addition, the chapter addressed the conceptual problem of using AI governance to analyse AI safety. It argued that 'AI governance' is a broad term, and that 'AI safety governance' is more suitable as a conceptual tool for discussing how a government governs AI development to ensure its safe and responsible use.

The current literature on AI safety tends to muddle AI concerns with unique characteristics under AI governance. The problem with this notion is that it limits our attentiveness to policy shifts that impact the future of advanced AI deployment. Investment in AI research and development is an aspect of AI governance; similarly, developing policies specifically to respond to AI safety issues is another form of AI governance. As a result, AI governance has become too broad a concept to describe, explain and understand policies that are particularly targeted at AI safety, to the advantage of deregulation of the AI sector.

The AI safety governance framework enables one to make sense of the policy shift reflected in Trump's (first and second terms) and Biden's administrations. When analysts focus their analysis on AI safety governance instead of the broader AI governance, government commitment to AI safety becomes clear. The framework also boosts policymakers' understanding of AI safety as a specific field under the broader AI governance framework. AI safety primarily focuses on policies that impact the safe and responsible development, deployment and decommissioning of AI models. The application of the framework enables AI governance professionals to identify their niche and equip themselves with knowledge of technical AI governance as a means of making sense of AI safety governance.

## References

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P. and Roberts, K. 2024. Artificial Intelligence risk management framework: Generative artificial intelligence profile (NIST Trustworthy and Responsible AI, NIST AI 600-1). *National Institute of Standards and Technology*. [Online]. Available at: https://doi.org/10.6028/NIST.AI.600-1 (Accessed 10 April 2025).

Batool, A., Zowghi, D. and Bano, M. 2025. AI governance: A systematic

literature review. *AI and Ethics*, 5: 3265—279. https://doi.org/10.1007/s43681-024-00653-w

Booth, H., Souppaya, M., Vassilev, A., Ogata, M., Stanley, M. and Scarfone, K. 2024. Secure software development practices for generative AI and dual-use foundation models: An SSDF community profile (NIST Special Publication 800-218A). *National Institute of Standards and Technology*. [Online]. Available at: https://doi.org/10.6028/NIST.SP.800-218A. (Accessed 12 April 2025).

Buhl, M. D., Sett, G., Koessler, L., Schuett, J. and Anderljung, M. 2024. Safety cases for frontier AI. [Online]. Available at: *https://arxiv.org/pdf/2410.21572*. (Accessed 20 April 2025)., Lee, Y., Fontana, P and Awad, G. 2024. Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency. *NIST*. [Online]. Available at: *https://www.nist.gov/publications/reducing-risks-posed-synthetic-content-overview-technical-approaches-digital-content* (Accessed 10 April 2025).

Delaney, O., Guest, O. and Williams, Z. 2024. Mapping technical safety research at AI companies: A literature review and incentives analysis. *Institute for AI Policy and Strategy*. [Online]. Available at: *https://arxiv.org/pdf/2409.07878*. (Accessed 10 April 2025).

Gruetzemacher, R.R. and Whittlestone, J. 2022. The transformative potential of artificial intelligence. *Futures,* 135. https://doi.org/10.1016/j.futures.2021.102884

.Hambling, D. 2024. Anduril's Bolt-M is the future for US forces, if they can afford it. *Forbes*. .Available at: *https://www.forbes.com/sites/davidhambling/2024/10/16/andurils-bolt-m-is-the-future-for-us-forces-if-they-can-afford-it/*. (Accessed 9 April 2025).

Han, T.A., Lenaerts, T., Santos, F.C. and Pereira, L.M. 2022. Voluntary safety commitments provide an escape from over-regulation in AI development. *Technology in Society*, 68. *https://doi.org/10.1016/j.techsoc.2021.101843*

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A. and Zilberman, N. 2024. *Governing through the cloud: The intermediary role of compute providers in AI regulation*. [Online]. Available at: *https://cdn.governance.ai/Governing-Through-the-Cloud_The-Intermediary-Role-of-*

*Compute-Providers-in-AI-Regulation.pdf* (Accessed 12 April 2025).

Hendrycks, D. 2024. *Introduction to AI Safety, Ethics, and Society*. [Online]. Available at: www.aisafetybook.com. (Accessed 10 April 2025).

Hendrycks, D., Mazeika, M. and Woodside, T. 2023. An overview of catastrophic AI risks. https://doi.org/10.48550/arXiv.2306.12001

Hilton, B., Buhl, M. D., Korbak, T. and Irving, G. 2025. Safety cases: A scalable approach to frontier AI safety. https://doi.org/10.48550/arXiv.2503.04744

Hubinger, E. 2025. Anthropic: Responsible scaling policy. *SuperIntelligence: Robotics: Safety & Alignment*, 2(1). *https://doi.org/10.70777/si.v2i1.13657*

Iqbal, S., Rizvi, S. W. A., Malik, M. H. H. and Raza, S. 2023. Artificial intelligence in security and defense: Explore the integration of AI in military strategies, security policies, and its implications for global power dynamics. *International Journal of Human and Society*, 3(4): 341-353.

Kramer, M. A., Leavens, A. and Scarlat, A. 2024. Harnessing AI for efficient analysis of complex policy documents: A case study of Executive Order 14110. .. Available at: *https://arxiv.org/pdf/2406.06657* . (Accessed:9 April, 2025).

Łucki, J., Wei, B., Huang, Y., Henderson, P., Tramèr, F. and Rando, J. 2024. An adversarial perspective on machine unlearning for AI safety. .. Available at: *https://arxiv.org/pdf/2409.18025*. (Accessed: 5 April, 2025).

Maini, P., Goyal, S., Sam, D., Robey, A., Savani, Y., Jiang, Y., Zou, A., Fredrikson, M., Lipton, Z.C. and Kolter, J. Z. 2025. Safety pretraining: Toward the next generation of safe ai. .. Available at: *https://arxiv.org/pdf/2504.16980*. (Accessed: 10 April 2025).

Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B. and Isaac, W. 2024. Generative AI misuse: A taxonomy of tactics and insights from real-world data. .. Available at: *https://arxiv.org/abs/2406.13843* (Accessed:5 April 2025).

.McGrath, A., Jonker, A. and Clark, B. 2024. What is AI safety? *IBM*. .. Available at: *https://www.ibm.com/think/topics/ai-safety*. (Accessed: 29 March, 2025).

*National Institute of Standards and Technology*. n.d. NIST AI Risk Management Framework Playbook. .. Available at: *https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook*. (Accessed: 29 March, 2025).

O'Brien, J., Ee, S. and Williams, Z. 2023. Deployment corrections: An incident response framework for frontier AI models. *Institute for AI Policy and Strategy*

*(IAPS)*. https://doi.org/10.48550/arXiv.2310.00328

Ramjee, D., & Selinger, E. (2025, May 2). How the U.S. Can Achieve Sustainable AI Leadership. *New Lines Institute for Strategy and Policy*. .. Available at:https://newlinesinstitute.org/wp-content/uploads/Cop-7-US-AI-Lead.pdf. (Accessed: 15 April, 2025).

Reuel, A., Soder, L., Bucknall, B. and Undheim, T. A. 2024. PositionTechnical research and talent is needed for effective AI governance. https://doi.org/10.48550/arXiv.2406.06987

Roberts, H., Hine, E., Taddeo, M. and Floridi, L. 2024. Global AI governance: Barriers and pathways forward. *International Affairs,* 100(3): 1275–1286. https://doi.org/10.1093/ia/iiae073

Roberts, H., Hine, E., Taddeo, M. and Floridi, L. 2024. Global AI governance: Barriers and pathways forward. *International Affairs,* 100(3): 1275–1286. https://doi.org/10.1093/ia/iiae073

Salvador, C. 2024. Certified safe: A schematic for approval regulation of frontier AI. https://doi.org/10.48550/arXiv.2408.06210

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. and Hall, P. 2022. Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.SP.1270

Shen, X., Chen, Z., Backes, M., Shen, Y. and Zhang, Y. 2024. 'Do anything now': Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City, Utah. pp. 1671—1685.

Szabadföldi, I. 2021. Artificial intelligence in military application–opportunities and challenges. *Land Forces Academy Review*, 26(2): 157—165. http://doi.org/*10.2478/raft-2021-0022*

Tabassi, E. 2023. Artificial intelligence risk management framework (AI RMF 1.0). *National Institute of Standards and Technology.* https://doi.org/10.6028/NIST.AI.100-1.

Tartaro, A., Smith, A. L. and Shaw, P. 2023. Assessing the impact of regulations and standards on innovation in the field of AI. .. Available at: *https://arxiv.org/pdf/2302.04110* (Accessed: 15 April 2025).

Tucker, P. 2025. What Google's return to defense AI means. *Defense One*. ..

Available at: *https://www.defenseone.com/business/2025/02/what-googles-return-defense-ai-means/402816/*. (Accessed: 15 April, 2025).

US AI Safety Institute. 2025. Managing Misuse Risk for Dual-Use Foundation Models (NIST AI-800-1 2pd, Second Public Draft). .. Available at: *https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf*. (Accessed: 15 April, 2025).

*US Department of State*. 2023. Political declaration on responsible military use of artificial intelligence and autonomy. .. Available at: https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/ (Accessed: 15 April, 2025).

United States. 2019. Executive Order 13859: Maintaining American leadership in artificial intelligence. *Federal Register*, 84(31): 3967–3969. .. Available at: *https://www.govinfo.gov/content/pkg/FR-2019-02-14/pdf/2019-02544.pdf*. (Accessed: 25 March, 2025).

United States. 2020. Executive Order 13960: Promoting the use of trustworthy artificial intelligence in the federal government. *Federal Register*, 85(236): 78939—78943.. Available at: *https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government*. (Accessed: 25 April, 2025).

United States. (Jan. 17, 2025). Executive Order 14141, Advancing United States leadership in artificial intelligence infrastructure. *Federal Register*: 5469—5489. .. Available at: https://www.federalregister.gov/documents/2025/01/17/2025-00765/advancing-united-states-leadership-in-artificial-intelligence-infrastructure. (Accessed: 15 April, 2025).

United States. 2023. Executive Order 14110: Safe, secure and trustworthy development and use of artificial intelligence. Federal Register: 75191—75226... Available at: https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. (Accessed: 25 April, 2025).

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X. and Hamin, M. 2025. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI-100-2 E2025). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.AI.100-2e2025