



# AI ELECTION RISKS AND US AI GOVERNANCE MEASURES

(AI に起因する選挙リスクと AI ガバナンス対策

米国調査レポート)

2024 年 12 月 独立行政法人情報処理推進機構

Prepared by Next Peak

## Table of Contents

1. Executive Summary.....	3
2. Index of Acronyms .....	5
3. Purpose, Scope, and Methodology .....	8
4. Background on AI Evolution.....	9
5. AI-Enabled Risks to Elections .....	11
A. Introduction.....	11
B. AI-Enabled Tactics for Misinformation and Disinformation .....	13
C. AI-Enabled Spear Phishing .....	17
D. AI-Enabled Disruptive Attacks.....	17
E. Common Targets of AI-Enabled Election Risks .....	17
F. Mitigation Methods.....	20
G. Summary.....	24
6. AI Governance Measures.....	26
A. Background.....	26
B. US AI Regulation and Framework.....	27
I. Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI.....	28
II. The Office of Science and Technology Policy (OSTP) Blueprint for AI Bill of Rights (AIBoR) ....	52
III. National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF): Generative AI Profiles.....	55
IV. Other NIST Drafts .....	57
V. Information technology - AI – Management system (ISO/IEC 42001:2023) .....	60
VI. Information technology - AI – Guidance on risk management (ISO/IEC 23894:2023) .....	62
VII. State Regulations .....	63
C. Summary.....	65
7. Criteria for Effective AI Governance and Framework Measures .....	68
A. Background.....	68
B. Findings from Research on US AI Governance Measures .....	69
I. Industry Self-Governance.....	69
II. AI Stacks .....	70
III. General-Purpose AI Systems.....	71
IV. AI Agency .....	71
V. Intellectual Property (IP).....	71
C. Findings from Expert Interviews.....	72
D. Approaches to Evaluating and Considering AI Governance Measures .....	74

I. MIT: A Framework for US AI Governance .....	74
II. WEF: Presidio AI Framework.....	75
III. CSET: Report on Flexible Approach.....	80
E. Criteria for AI Governance Approaches.....	84
I. Pillar 1: Foundational Principles.....	85
II. Pillar 2: Higher-Level Strategies .....	85
III. Pillar 3: Sector-Specific Prescriptive Regulations .....	86
F. Assessment of Current US AI Governance Measures.....	87
8. Conclusion.....	91
9. Expert Interviews .....	93
AI Policy Expert 1 .....	93
Security Expert 1 .....	96
AI Policy Expert 2 .....	98
Policy Expert 1.....	103
AI Policy Expert 3 .....	106
AI Policy Expert 4 .....	109
Government Agency 1 .....	111
Election Analyst 1.....	115
AI Policy Expert 5 .....	118
Data Consultant 1 .....	121
10. Appendix.....	126
Appendix A: Case Studies of AI in Recent Elections.....	126
Appendix B: Overview of Recent Global AI Regulation.....	132
Appendix C: Comparison Between the NIST AI RMF and Japan’s AI GfB.....	134
Appendix D: Table of State-Level AI Regulation .....	135
Appendix E: Strengths and Weaknesses of Select US AI Governance Approach.....	138
11. Annotated Bibliography.....	140
12. References.....	142
13. Footnotes.....	153

## 1. Executive Summary

Since the mainstream emergence and democratization of artificial intelligence (AI) technologies in late 2022, generative AI (GenAI) tools—including generative text, audio, image, and video—have become critical to industry operations, accessible to the wider population, and beneficial to cyber threat actors.<sup>1</sup> Furthermore, discussions around GenAI regulations have gained traction throughout 2023 as the technology became more powerful and pervasive.<sup>2</sup>

The first report that Next Peak prepared for the Information Protection Agency (IPA) in 2024 focused on five areas of AI-enabled risk: 1) AI-enhanced traditional cyberattacks; 2) AI-enabled disinformation; 3) AI-enabled disruption or mis-operation of systems; 4) AI-enabled national security threats and 5) business risks due to misuse of GenAI. This second report builds on the first report by conducting a survey of over 190 sources and 10 expert interviews to investigate the following two topic areas.

The first topic area is AI-enabled election risks. As nearly half of the world voted or will vote in 2024,<sup>3</sup> pre-existing election risks including misinformation, disinformation, disruptive attacks, and phishing attempts were magnified by AI technologies.

Around the world, Indonesia saw a proliferation of AI-generated deepfakes and misinformation, largely due to the high penetration of social media in the country as well as the Indonesian General Election Commission's lack of regulations. AI-generated deepfake audio also interfered with the Pakistani elections as a clip of Imran Khan suggesting an election boycott circulated. Foreign adversaries—Russia and China—also used AI to interfere in other elections: China conducted an AI-enabled disinformation campaign to influence the Taiwanese election; Russia used AI-generated deepfakes and false information in the Slovakian, United States (US), United Kingdom (UK), and Moldovan elections.

The 2024 US presidential election is on November 5<sup>th</sup>, and the US has also been seeing an increase in AI-enabled election interference—on social media platforms such as X, Telegram, and WhatsApp—as the election date approached. The first significant case was the 2024 AI-generated robocalls of US President Biden discouraging voting in the New Hampshire Primary. There were various AI-generated deepfakes, misinformation from AI chatbots regarding election news and updates, AI-enabled foreign interference campaigns by Iran, North Korea, Russia, and China, and an AI-generated malware—AsyncRAT—microtargeting specific communities.

## A year of elections

Over two dozen economies are scheduled to hold general elections in 2024.

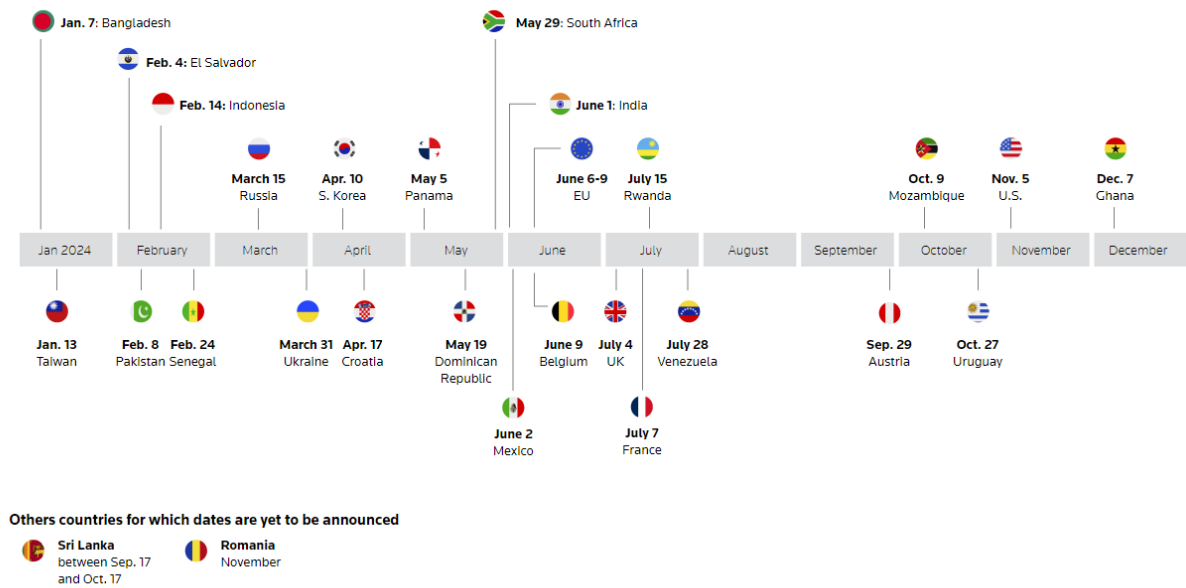


Figure 1: Timeline of 2024 Elections<sup>4</sup>

The second topic area is AI governance measures in the United States. From international AI governance measures such as the European Union's AI Act to the US President Biden's AI Executive Order (EO) to various state-level AI initiatives, there has been an increase in inconsistent and varying AI regulations: around 27% of Fortune 500 companies cited AI regulation as a risk in their recent filings with the Securities and Exchange Commission, demonstrating the various risks that come with an explosion in the regulatory space.<sup>5</sup> This report focuses specifically on US AI governance measures and framework.

At a high level, no single US governance measure is comprehensive of all risks, but in combination, the US collectively cover the five AI-enabled risks covered in Report 1: pre-existing cybersecurity regulations and the AI EO covered AI-enhanced traditional cyberattacks. The AI Bill of Rights as well as additional acts combatted AI-enabled disinformation such as deepfakes. The AI EO and the AI Bill of Rights focused on safe and effective systems, covering AI-enabled disruptions and maloperations of systems. The NIST RMF and the NIST Plan for Global Engagement on AI Standards strive to protect against AI-enabled national security threats. Finally, the NIST SSDF for GenAI focused on secure development practices and dual-use foundation models, mitigating against AI-enabled business risks.

Based on the research and investigation, the report proposes an approach to evaluating existing AI governance frameworks and provides key principles and factors to consider for effective AI governance measures. Finally, the evaluation criteria are applied to current US AI governance measures to determine the strengths and weaknesses of each regulation.

## 2. Index of Acronyms

ADA	Americans with Disabilities Act
ADMT	Automated Decision-Making Tools
AEDT	Automated Employment Decision Tool
AI GfB	Japan AI Guidelines for Business
AI	Artificial Intelligence
AIBoR	Blueprint for AI Bill of Rights
AIDA	European Union (EU)'s AI Act and Canada's AI and Data Act
AIMS	AI Management Systems
AISSB	AI Safety and Security Advisory Board
API	Application Programming Interface
ARIA	NIST's Assessing Risks and Impacts of AI program
ATLAS	Adversarial Threat Landscape for AI systems
B2B	Business-to-Business
B2C	Business-to-Consumer
BIS	Bureau of Industry and Security
BJP Party	Bharatiya Janata Party of India
BSA	The Software Alliance
CAI	Commercially Available Information
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CBRN	Chemical, Biological, Radiological, and Nuclear
CCPA	California Consumer Privacy Act
CFPB	Consumer Financial Protection Bureau
CIP	Customer Identification Program
CISA	Cybersecurity and Infrastructure Security Agency
CISO	Chief Information Security Officer
CPRA	California Privacy Rights Act of 2020
CSAM	Child Sexual Abuse Material
CSET	Centre for Security and Emerging Technology
CWMD	Countering Weapons of Mass Destruction
DDoS)	Denial-of-Service
DFPI	Department of Financial Protection and Innovation
DFS	Department of Financial Services
DHS	Department of Homeland Security
DIT	Department for International Trade
DKIM	DomainKeys Identified Mail
DMARC	Domain-based Message Authentication, Reporting and Conformance
DOC	Department of Commerce
DOD	Department of Defense
DOE	Department of Energy
DOL	Department of Labor
DoS	Department of State

DPA	Defense Production Act
EDR	Endpoint Detections and Response
EEOC	Equal Employment Opportunity Commission
EI-ISAC	Center for Internet Security's Elections Infrastructure Information Sharing and Analysis Center
EO	Executive Order
EOP	Executive Office of the President
EU	European Union
FDA	Food and Drug Administration
FINRA	Financial Industry Regulatory Authority
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
HAI	Stanford Institute for Human-Centered Artificial Intelligence
HHS	Department of Health and Human Services
HUD	Department of Housing and Urban Development
IaaS	Infrastructure as a Service
IC	Intelligence Community
IDB	Inter-American Development Bank
IEC	International Electrotechnical Commission
IoT	Internet of Things
IPA	Information Protection Agency
ISO	International Organization for Standardization
IT	Information Technology
J-AISI	Japan AI Safety Institute
JCDC	Joint Cyber Defense Collaborative
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Models
MFA	Multifactor Authentication
NAACP	National Association for the Advancement of Colored People
NAIRR	National AI Research Resource
NCII	Non-Consensual Intimate Imagery
NGOs	Nongovernmental Organizations
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NSF	National Science Foundation
NTIA	National Telecommunications and Information Administration
ODNI	Office of the Director of National Intelligence
OECD	Organization for Economic Co-operation and Development
OMB	Office of Personnel Management
OPM	Office of Personnel Management
OSTP	The White House Office of Science and Technology Policy
OWASP	Open Worldwide Application Security Project
PET	Privacy-Enhancing Technologies
PII	Personally Identifiable Information

PO	Preparing Organization
PS	Protecting Software
PW	Producing Well-Secured Software
RBRs	Rule-Based Rewards
RCN	Research Coordination Network
RFI	Request for Information
RLAIF	Reinforcement Learning from AI Feedback
RLHF	Reinforcement Learning from Human Feedback
RMF	AI Risk Management Framework
RV	Responding to Vulnerabilities
SDLC	Software Development Life Cycle
SEC	Securities and Exchange Commission
SHAP	SHapley Additive exPlanations
SPF	Sender Policy Framework
SRMAs	Sector Risk Management Agencies
SSDF	NIST Secure Software Development Framework
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
USAID	United States Agency for International Development
US	United States
USPTO	United States Patent and Trademark Office
VA	Department of Veterans Affairs
VR	Virtual Reality
WEF	World Economic Forum
XAI	Explainable AI



### 3. Purpose, Scope, and Methodology

Even before 2024 started, experts at the Stanford Law School and MIT Media, “many polls...[showed] just how high the anxiety is in the general public about the impact of artificial intelligence on our elections,” and even “AI panic is itself a democracy problem.”<sup>6</sup> While AI-enabled risks towards all critical infrastructure is threatening, AI-enabled election risks have high visibility, especially in 2024. Thus, this report analyzes cyber threats posed by AI in elections and assesses the efficacy of AI governance protocols in mitigating such risks.

This report seeks to inform the IPA Security Center and the Japan AI Safety Institute (J-AISI) established within the IPA in February 2024. J-AISI strives to study evaluation methods for AI safety and will act as a counterpart to US AISI within the National Institute of Standards and Technology (NIST).<sup>7</sup> AI risks are extensive and expand across various industries and digital domains, and the breadth of AI threats and risks will be provided for context and based on the first report Next Peak prepared for the IPA in 2024. The report will focus on AI risks and their impacts on elections, case studies of AI-enabled election risks seen thus far, mitigation methods, and positive election-related AI use cases. Then, there will be a detailed analysis of significant US governance and framework measures for AI and propose criteria for identifying effective AI regulations and strategies. Finally, the criteria will help determine the strengths and weaknesses of US AI governance measures, and the report will conclude with recommendations for further AI initiatives. This report was produced in multiple phases:

1. First, an extensive literature review of AI-enabled election risks, US approaches to regulating AI, and evaluation of AI governance measures was conducted. The review included research, publications, and articles about potential AI-enabled election risks as well as case studies. Drawing on previous research and the existing literature, the report considers how AI-enabled election risks could be mitigated. In the literature review, over 190 individual sources were surveyed. The annotated bibliography highlighting 10 significant sources is included as well.
2. Next, previous research, current literature review, and existing interviewee base were scoped to identify a list of AI and election experts to engage. The process informed the development of key interview questions tailored to address these critical issues. Subsequently, interviews with a diverse selection of experts from across the public, private, and non-profit sectors were conducted. The interviewees’ expertise and perspectives were also wide-ranging, covering technical AI development to legal and policy frameworks to election processes. The individual interviews and interviewee bios are included in the Expert Interviews section.
3. Key findings and insights were synthesized from the research, literature review, and interviews to propose three-phase criteria for AI governance measures.

## 4. Background on AI Evolution

The complexity of AI technology is constantly transforming, leading to a surge in capabilities and applications over the last few years. AI models and systems are difficult to categorize, which adds to the layer of complexity in AI-enabled risks as well. The table below—created in April 2024 for the first Next Peak Report for the IPA—summarizes the current threat landscape due to AI and attempts to estimate the timeline of risks and severity of threats.

Table 1: Summary of AI Threats and Risk Chart

Threat	Risk	Impacted sector/entity/etc.	Timeline <sup>i</sup>	Impact
AI-enhanced traditional cyberattacks	Force multiplier for disruptive attacks	All sectors but critical infrastructure may be impacted greatly	Medium term	High
	Increased capabilities and efficiency of cybercriminals in ransomware and cryptocurrency-related cyberattacks; lowered barrier to entry	Individuals and industries, especially ransomware-prone ones such as health care, financial, and hospitality	Medium term	High
	Lowered barrier to entry for social engineering; increased efficiency and speed in spear phishing	Individuals, industries, governments, academia, news organizations, critical infrastructure	Immediate	High
AI-enabled disinformation	Domestic Disinformation: increased censorship, targeting of vulnerable groups, spread of authoritarian digital norms	Particularly individuals and minorities in authoritarian nations, democracy, freedom of speech	Immediate	Medium
	State-sponsored disinformation campaigns: polarization of societies, erosion of trust, degrading of democracy	Individuals, democratic governments, electoral process Democratic	Immediate	Medium
	Promotion of crime and discrimination: new class of crime such as deepfake pornography and stock market manipulation	Individuals, finance industry, black market, private sector widely	Medium term	Medium
	Election Obstruction: online censorship, disinformation	Individuals, freedom of speech, democratic nations, electoral process	Immediate	Medium-High
AI-Enabled disruption or maloperation of systems	Data poisoning: false outputs leading to bad decision-making, discrimination, disruption	Critical infrastructure, social infrastructure, justice system, others	Medium term	High
	Inherent biases and vulnerabilities: reinforce stereotypes, biased content generation and decision-making	Individuals, businesses, governments	Immediate	Medium

<sup>i</sup> Immediate refers to happening currently or in the next couple of years. Medium-term refers to the next 3-5 years. Long-term refers to the next 5-10 years.

Threat	Risk	Impacted sector/entity/etc.	Timeline <sup>i</sup>	Impact
	Intentional and unintentional failures: operational disruption and false outputs	Critical infrastructure, social infrastructure, justice system, multiple industries	Immediate	Medium-High
AI-enabled national security threats	Military applications: potential autonomous weapon systems, military decision making leading to ethical concerns	Defense sector, governments	Long term <sup>8</sup>	High
	AI race: deployment of AI systems with unproven reliability, risk of escalation	Governments, defense sector, industry	Long term	High
	Espionage and Mass Surveillance: higher scale and speed, erroneous uses by the private sector	Public and private sector, individuals, privacy	Medium term	Medium
	Terrorism: dissemination of propaganda, assist with terrorist plans	Social media companies, individuals, governments	Medium term	Low
	Bioterrorism: development of novel pathogens, efficient information gathering	Individuals, healthcare, and pharmaceutical sectors	Long term	Low
Business risks due to misuse of GenAI	Vulnerable code generation and dissemination (can be due to insufficient oversight and testing): data leakage, reputational damage, regulatory noncompliance, financial losses, operational disruption	Businesses, consumers, employees, privacy	Immediate	Medium
	Legal risks and insider threats: data leakage, trade secret theft, noncompliance, financial penalties	Legal system, privacy, businesses, individuals	Immediate	Medium

The risks above apply to election risks as well from AI-enabled disinformation that erodes the public's trust in election processes and results to AI-enabled disruptive attacks that target election infrastructure.

## 5. AI-Enabled Risks to Elections

### A. Introduction

Experts warn of the growing risks due to AI technologies towards elections as threat actors and political campaigns seek to leverage AI to influence the electoral process, from spreading disinformation to manipulating voting systems.<sup>9</sup> Despite social media and news' depiction of GenAI as creating various new and complex challenges to civil and political society, experts tend to agree that GenAI will not present completely new threats to elections. Likely, the proliferation of GenAI will augment existing risks and issues.<sup>ii</sup>

The evolution of election threats has become increasingly complex with the rise of AI technologies, which affect nearly every stage and stakeholder of the election process. Though election processes vary across democracies, certain core components remain consistent—such as voter registration, announcements and campaigns from candidates, the process of voting itself, tallying of votes, certification of results, and transition of power—placing AI-enabled election risks on all democracies. Particularly in the US, non-centralized election processes—in which each state administrates and manages its own electoral process and voting day logistics—contribute to a level of protection. Centralization of electoral processes would allow for one risk or threat to cascade into a domino effect that leads to the collapse of the singular electoral system.<sup>iii</sup>

Over the past decade, the US information environment has undergone three phases—due to the evolution of the political environment, news cycles, culture, and technologies—that correlated with the US presidential election cycle. These three phases can generally be identified as the 2016 Cambridge Analytica Era, the 2020 Twitter Era, and the 2024 GenAI Era. An analysis of these three eras demonstrates that election threats such as foreign interference, disinformation, or information warfare have existed since before 2016. As experts warn, these preexisting threats are increasing as threat actors and political campaigns seek to leverage AI to influence the electoral process in more sophisticated ways.

It is important to note that the past two US presidential elections and the upcoming one in November 2024 have been plagued with additional tumultuous factors that augment election risks. The 2016 election (Clinton v. Trump) was only the fifth time in which a candidate who won the popular vote did not win the electoral vote and the election.<sup>10</sup> The 2020 election (Biden v. Trump) was significantly influenced by the coronavirus pandemic.<sup>11</sup> The upcoming 2024 presidential election (Harris v. Trump)<sup>12</sup> had a rocky start with Biden making a historic decision to drop out of the race.<sup>13</sup> Within the US context, the following table characterizes the three eras and includes election risk examples.

---

<sup>ii</sup> Election Analyst 1 interview.

<sup>iii</sup> Security Expert 1 interview.

Table 2: Evolution of AI-enabled threats to US elections, 2016-2024

Era	Description	Examples
Cambridge Analytica Era 2016	<ul style="list-style-type: none"> <li>Characterized by foreign entities targeting election campaigns and election infrastructure, including voter registration databases.</li> <li>The US Senate Intelligence Committee reported that the Russian government orchestrated voter data collection and used troll farms/bots to spread false information about elections, election candidates, and hot-topic issues in the 2016 US election.</li> <li>Russia also organized competing rallies, sometimes employing Americans, over polarizing issues including racial politics. Groups with opposing views converged and fostered anger and discord. As elections approached, attention shifted to utilizing augmenting emotional reactions through news articles and other media coverage.<sup>iv</sup></li> </ul>	<ul style="list-style-type: none"> <li>Globally, Russia allegedly interfered with UK’s decision to pass Brexit.<sup>14</sup></li> <li>The use of data targeting and microtargeting by Cambridge Analytica to develop and distribute data-driven services to various political campaigns.<sup>15</sup></li> </ul>
Twitter Era 2020	<ul style="list-style-type: none"> <li>Cyber activities persisted, but tactics shifted to sowing doubts about the integrity of electoral processes, including fraud, ballot manipulation, and Trump’s claims of a stolen election.</li> <li>Tactics shifted to a bottom-up approach: smaller social media accounts created narratives that were amplified by larger accounts of prominent domestic and foreign influencers, which presented significant challenges to social media companies in terms of combatting mis- and disinformation.</li> <li>Information warfare shifted from the creation of new stories to the augmentation of existing narratives. By capitalizing on already-existing narratives, existing tensions were exacerbated leading to pollution of the information space.<sup>v</sup></li> </ul>	<ul style="list-style-type: none"> <li>Claims of election fraud and manipulation gained traction on social media as influencers and large sources amplified content from smaller accounts. Some cyber and AI-related issues persisted, but greater concerns arose from general distrust of the veracity of election information.<sup>vi</sup></li> <li>False stories of a poll worker in Pennsylvania discarding ballots gained traction.<sup>16</sup></li> </ul>
GenAI Era 2024	<ul style="list-style-type: none"> <li>In the US context, emerging legal norms and policy discussions have curtailed use of GenAI by campaigns. Increasing incidents of existing threats from other actors have emerged, including phishing attacks on election officials, robocalls or messages that adapt to user input, and cyberattacks on infrastructure. Attacks emerge from both domestic and foreign threat actors.</li> <li>Robocalls, phishing incidents, deepfakes, disinformation campaigns, and other incidents are increasingly targeted at specific populations.<sup>vii</sup></li> </ul>	<ul style="list-style-type: none"> <li>Social media accounts pushed inaccurate stories that claimed President Joe Biden had died.<sup>17</sup></li> <li>AI-generated deepfakes featuring Bollywood actors making claims about political parties went viral and were subsequently denounced as fake.<sup>18</sup></li> </ul>

<sup>iv</sup> Policy Expert1 interview.

<sup>v</sup> Policy Expert 1 interview.

<sup>vi</sup> Policy Expert 1 interview.

<sup>vii</sup> Election Analyst 1 interview.

As prefaced in the table above, the rapid adoption of AI technologies across industries, academia, and civil society represents a myriad of risks to the proper functioning of electoral systems and confidence in the outcomes of elections. Election risks are both complex and broad, ranging from AI chatbots delivering misinformation to AI-generated disinformation via deepfake videos and audios to phishing campaigns that target voters and candidates to cyberattacks on election infrastructure. However, the emergence of completely novel risks is unlikely to present significant problems; rather, it is more likely that AI technologies will exacerbate existing vulnerabilities and weaknesses within election systems and processes. For example, misinformation and disinformation have always been election threats, and AI is exacerbating this existing threat with AI-generated deepfakes. Additionally, it is important to note that experts lack consensus on the impact of AI technology.<sup>19</sup>

As AI-enabled risks target the different layers and dimensions of elections, from the election infrastructure to the media to the voters,<sup>viii</sup> it is essential that governments and other organizations stay at the vanguard of addressing novel and pre-existing threats to local and national elections. Many of these threats can be addressed by enhancing pre-existing cybersecurity best practices and tactics. Governments have thus far responded with urgency through the introduction of regulations and guidance: from the Biden administrative directive encouraging over 50 federal entities to incorporate specific requirements in their policies to Taiwan's legislative amendments that introduce penalties for the creation and distribution of deepfake video and audio.<sup>20</sup> Additionally, traditional methods to combat cyberattacks can be leveraged to defend against AI-enhanced efforts, and AI-specific mitigation methods are also starting to surface.

The following section will delineate election threats, first focusing on AI-enabled amplification of disinformation and misinformation operations and then on other attacks that can target election processes, offices, officials, and vendors. The section will also include case studies of recently seen AI-enhanced election risks and conclude with various mitigation methods to safeguard democracy as well as free and fair elections.

## B. AI-Enabled Tactics for Misinformation and Disinformation

Both foreign and domestic actors can conduct misinformation and disinformation campaigns. Though tactics of both domestic actors and foreign actors overlap, processes, objectives and targets may differ.

Foreign malign influence, or the coordinated actions by foreign actors to manipulate opinions, behaviors, or decisions within a target country has been utilized for decades. From propaganda to disinformation utilized to directly impact public opinion, existing threats face

---

<sup>viii</sup> Security Expert 1 interview.

exacerbated effects from AI technologies, particularly large language models (LLMs),<sup>21</sup> that enable broader reach, cheaper initiation, and more convincing narratives.<sup>22</sup> The figure below provides an outline and further detail of the tactics detailed above.

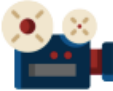





TACTIC	EXAMPLE
 <p><b>DISGUIISING PROXY MEDIA</b> Foreign malign influence actors disguise proxy media used to spread content as established media, by impersonating established outlets and local news sources.</p>	<p><i>PRC actors used AI news anchors for fictitious media outlets to spread pro-PRC content.<sup>xv</sup></i></p>
 <p><b>VOICE CLONING PUBLIC FIGURES</b> A fabricated recording of a public official is used to mislead the public or a targeted individual.</p>	<p><i>Voice clones of a political party leader were disseminated in the Slovak Republic two days before the election as reported by Wired magazine in October 2023.<sup>xvi</sup></i></p>
 <p><b>CYBER-ENABLED INFORMATION OPERATIONS</b> Foreign adversaries compromise IT systems of prominent organizations to find and leak damaging private information.</p>	<p><i>Russian Federation actors hacked and leaked US-UK trade documents prior to Britain's 2019 election according to Reuters.<sup>xvii</sup></i></p>
 <p><b>MANUFACTURING FALSE EVIDENCE OF SECURITY INCIDENT</b> A false report of a physical or cybersecurity incident is spread.</p>	<p><i>Pro-PRC actors distributed fake leaked Taiwanese government documents before the Taiwanese elections.<sup>xviii</sup></i></p>
 <p><b>PAID INFLUENCE</b> Foreign malign influence actors launder messaging by covertly paying online influencers, hiring PR firms, or employing journalists to spread disinformation.</p>	<p><i>The PRC use influencers to push foreign malign influence content about Xinjiang on Western social media.<sup>xix</sup></i></p>
 <p><b>LEVERAGING SOCIAL MEDIA PLATFORMS</b> Foreign nation-state actors leverage social media platforms to spread influence narratives in specific communities.</p>	<p><i>Russian Federation actors leveraged social media platforms with less stringent content-moderation policies to spread divisive political narratives in the U.S.<sup>xx</sup></i></p>

Figure 2: Foreign Malign Influence Tactics<sup>23</sup>

Additionally, the following operations are employed by both domestic and foreign actors aiming to foment mistrust, spread false narratives, or to influence and shape policy and discourse among targeted governments, organizations, and other actors. In the context of elections, each operation poses unique risks to the information space, which can be further augmented by using GenAI tools to create more widespread reach and increase the believability of operations.<sup>24</sup>

## **Proxy Media and Misleading Personas**

Websites or networks of fake personas are produced with the intent to appear independent and with the aim of suggesting authenticity. Leveraging this false credibility, threat actors disseminate false information. Tactics to increase credibility include imitating trusted sources like think tanks, journalists, or professors.<sup>25</sup> In 2022, experts identified instances of a pro-Chinese influence operation utilizing content depicting fictitious people as almost certainly originating from GenAI technologies. Though content was fairly low quality and seen by few, commercially available AI tools now enable the creation of higher-quality content which can be spread at a much faster rate than in the past.<sup>26</sup>

## **Voice Cloning and Deepfakes**

Malicious actors may create manipulated or entirely fake videos, images, and audio. AI technologies are increasingly able to create content that is nearly indistinguishable from authentic content, enabling more efficient manipulation of audiences. For example, in the Slovakian election in 2023 a deepfake recording circulated of the progressive party leader making statements in line with pro-Russia narratives. Party leadership ultimately called the video and statement “made up.”<sup>27</sup> Deepfakes make it increasingly difficult to distinguish fact from fiction due to the widespread wariness of false information, leading to the threat of the “Liar’s Dividend.”<sup>28</sup>

## **Conspiracy Theories, Information Operations, Manufactured Evidence**

Conspiracy theories seek to explain important events by attributing their cause to secret plots or actions devised by powerful actors. Such theories can affect the audience’s view of the event in question but also their worldview overall. Disinformation campaigns and misinformation can play into this, creating stories that resonate with particular audiences and further exacerbating that audience’s belief in false narratives.<sup>29</sup> For example, After President Biden withdrew from the 2024 US presidential election, there were rumors circulating on X that President Biden had passed away. X recommended the post, and its AI software summarized this rumor as a trending story.<sup>30</sup>

Information operations utilizing cyber intrusions—when a threat actor hacks or compromises systems of prominent organizations and subsequently distribute private or sensitive information, sometimes framing the organization or others as having leaked the information—also poses risks such as reputational damage.<sup>31</sup>

In recent years Iran has engaged in several different initiatives to sow discord in Israel such as compromising IT systems tied to the Israeli government. Then, hackers publicized these incidents as cyberattacks undertaken by domestic Israeli activists rather than Iran. As of



October 2024, there has yet to be a use case of AI-enabled cyber intrusion combined with an information operation. In the future, AI can be used to create malware that adapts more readily to organizations' security systems, attacks like the above may become easier to initiate.<sup>32</sup>

In the future, false evidence of events that have not actually taken place may also emerge. For example, fake cybercriminal personas may be leveraged to spread hacked documents or false reports, which includes creating false records of security incidents.<sup>33</sup>

### **Paid Influence and Leveraging Social Media**

Both foreign and domestic actors may pay influential people or organizations to push their messaging for the purpose of lending false narratives more credibility. Such influence may be leveraged through the employment of internet entities and influencers as well as public relations firms.<sup>34</sup> Content created to be easily shareable—in the form of videos, post or memes—may also exploit paid advertising models to gain a wider audience.<sup>35</sup> AI can rapidly produce and disseminate such information, acting as a force multiplier for disinformation campaigns.<sup>36</sup>

AI-enabled Bots or groups of actors known as “troll farms” can disseminate heavy volumes of false narratives from inauthentic accounts. The sheer volume appears as wide grassroots support for a particular point of view that is inauthentic, and this tactic is known as “astroturfing”.<sup>37</sup> Similarly, accounts may spam social media, forums, and comment sections with the intention of flooding out legitimate conversation. This sort of spamming discourages legitimate actors' participation while overwhelming targets to the extent that they no longer readily believe in the truth of information presented to them.<sup>38</sup>

### **Microtargeting**

Foreign and domestic malicious actors may utilize social media to disseminate false narratives and manufactured stories that are highly tailored. This tactic, known as microtargeting, aims to influence specific communities and to give threat actors “insider status” with highly believable content.<sup>39</sup> AI can help tailor or exacerbate narratives for microtargeting and polarization.<sup>40</sup>

Experts are increasingly concerned about AI-enhanced microtargeting in elections. Since GenAI enables the dissemination of information to receptive audiences at a faster rate and to a wider degree, microtargeting of certain population groups with specific disinformation

narratives and messages can happen at an increased speed, scale and scope.<sup>ix</sup> Campaigns can be customized to exploit voter's racial, ethnic, religious, or other identities.<sup>41</sup>

### C. AI-Enabled Spear Phishing

AI enables increased personalization and sophistication of phishing campaigns which lead to a proliferation of attacks. Research indicates that users fall victim to AI-enabled spear phishing campaigns at about the same rate as those written by humans, but AI enables significant reductions in costs allowing rapid growth of attacks.<sup>42</sup> Moreover, while falsified text, voice, and videos created by GenAI technologies make it easier and cheaper to create spear phishing attacks, preventative measures including education remain costly.<sup>43</sup> For example, SweetSpecter, a suspected China-based adversary, used ChatGPT to enhance spear phishing attacks against OpenAI employees and governments in 2024. In the context of elections, AI-enhanced phishing attacks may utilize more convincing communications to trick voters, candidates, and election officials.<sup>44</sup>

### D. AI-Enabled Disruptive Attacks

Distributed denial-of-service (DDoS) attacks aim to flood and incapacitate systems with mass amounts of information. With free tools to facilitate such attacks readily available online and AI further augmenting the ability of such attacks to wreak havoc, challenges remain to defend against DDoS attacks. Malicious actors may utilize the Internet of Things (IoT) to create a botnet and multiply attacks and leveraging more sophisticated AI technologies can further proliferate attacks. Both malicious actors and defenders can use AI to predict and combat attacks and defenses, leading to possibly larger attacks that can engulf systems or services.<sup>45</sup> Such attacks may render access to election information and services difficult for voters, including voter registration or unofficial election results websites.<sup>46</sup>

AI may also be leveraged to augment other types of disruptive attacks. Threat actors can craft advanced malware that adapts its behavior to past experience and is undetectable by endpoint detections and response (EDR) applications<sup>47</sup> which may affect election infrastructure.<sup>48</sup> Though experts fear the use of large language models (LLMs) to find and exploit vulnerabilities or create malicious code, evidence suggests that such technologies have more frequently been utilized to translate documents, draft emails, or debug code.<sup>49</sup>

### E. Common Targets of AI-Enabled Election Risks

AI-enabled cyber threats to elections generally center on targets broadly defined within the following five categories: 1) electorate 2) election processes; 3) election offices; 4) election

---

<sup>ix</sup> Election Analyst 1 interview.

officials; 5) election vendors. Previously, sections 4.B and 4.C of this report detail the various potential AI-enabled misinformation, disinformation, and spear phishing attacks that could target the electorate, sow discord, further polarize voters, and manipulate voter sentiment. Recent uses of AI-generated images aimed at the US have sought to reduce public support for providing military and financial aid to allies while deepening divisions along racial, economic, and ideological divides.<sup>50</sup> In response to these emerging threats, CISA issues guidance outlining how malicious actors might use GenAI capabilities to influence elections.

The figure below outlines the potential risks of GenAI to distinct election-related targets beyond the electorate.

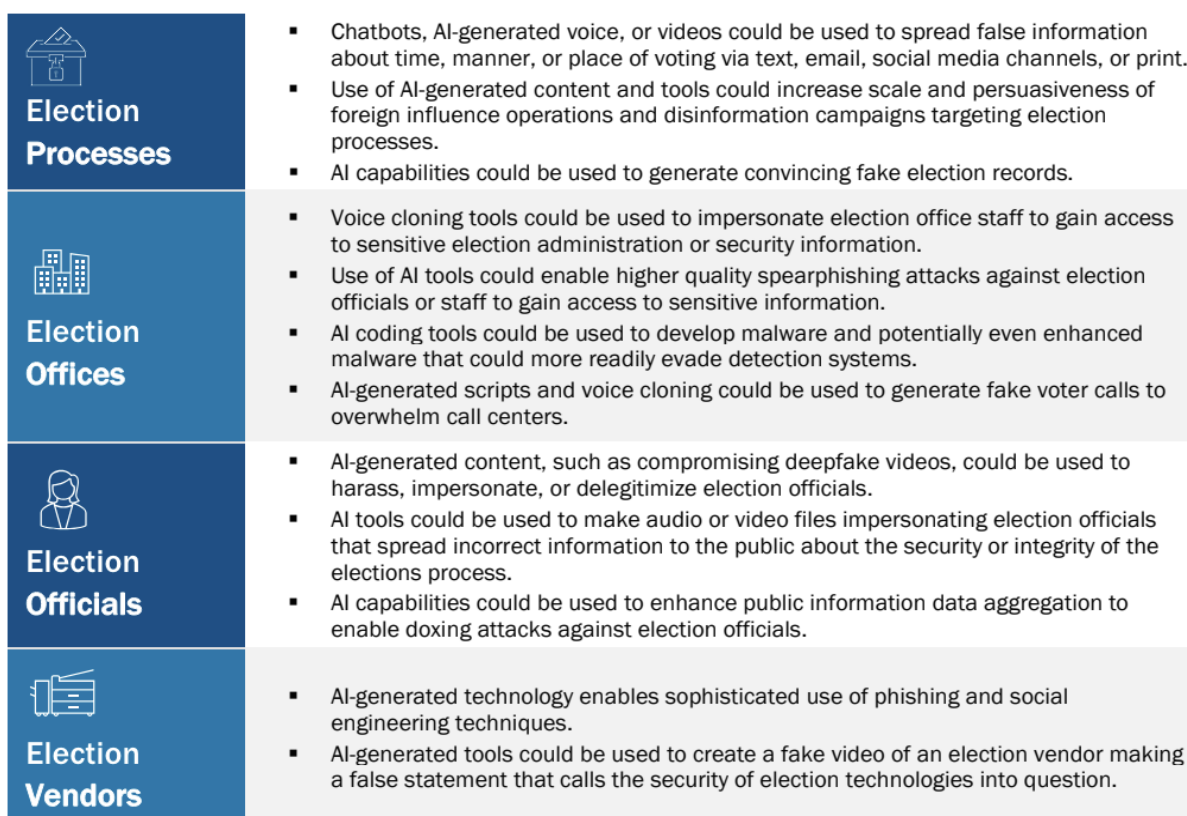


Figure 3: Malicious uses of GenAI in the context of elections<sup>51</sup>

### Election Processes

Election processes encompass all that help run elections smoothly. The dissemination of correct information about the location, timing, and logistics of voting is critical<sup>52</sup> and now relies on the internet, social media, and electronic communications. Election infrastructure, such as voting equipment that moves votes and reports to servers as well as voter databases, is also dependent on various networks, systems, the cloud, computers, and software.<sup>x</sup> Thus, the election process is already susceptible to disruptions caused by malicious actors using

<sup>x</sup> Security Expert 1 interview.

malware, phishing, and more as well as issues caused by using older election technologies. Threat actors are likely to use GenAI to further spread disinformation about election logistics, to contribute to increased scale and impact of foreign influence operations, and to produce fake voter records.<sup>53</sup> For example, earlier this year a robocall created by AI technologies imitating President Biden’s voice downplayed the importance of voters participating in primary elections in New Hampshire.<sup>54</sup>

### **Election Offices**

Offices may face issues including the malicious or negligent access of confidential voting information by staff or deliberate campaigns that overwhelm the communications capabilities and systems of offices providing trouble-seeking and guidance during every step of the election process from registration to vote counting. AI voice cloning tools may be utilized to impersonate staff to access security information, threat actors may engage spear-phishing attacks against officials and staff to access sensitive information, AI may be used to create new versions of malware that evade detection, and AI-generated scripts and voice cloning tools may be leveraged to overwhelm call centers, among other threats.<sup>55</sup> For example, in 2022 the Mississippi Secretary of State’s website suffered a DDoS attack on election day preventing public access to information.<sup>56</sup>

### **Election Officials**

Officials staffing all levels of election processes face myriad AI threats, spanning from impersonation to doxing to physical safety threats. AI-generated content, including voice cloning tools and deepfakes, can be used to harass, impersonate, or delegitimize officials and candidates. Such technologies can also be utilized by threat actors to spread misinformation about election processes and procedures, or to call into question the legitimacy or integrity of the election overall. Moreover, generative technologies can enable further ease of aggregating and disseminating personal information, possibly increasing doxing incidents for officials, staff, and candidates. For example, in the leadup to the 2024 US presidential election, phishing emails impersonating the US Election Assistance Commission (EAC)—a body of election officials—prompted recipients to review and confirm voter registration. The US EAC does not track or store voters’ personal information, and the malicious actors were likely targeting voter information.<sup>57</sup> If AI were to be used in such attacks, threat actors will be able to better impersonate election officials.

### **Election Vendors**

Election vendors—such as companies producing electronic voting machines—are companies contracted by the government or election bodies to provide infrastructure and processes for elections.<sup>58</sup> AI technology utilized to question the legitimacy and security of election vendors

may augment risks such as general chaos to election stakeholders and disruption of genuine engagement in the election process. Social engineering techniques in addition to deepfakes can allow impersonation of vendors, undermining voting apparatuses' integrity.<sup>59</sup> In 2016 Russian hackers sent phishing emails to employees at electronic voting software vendor VR Systems, redirecting them to fake websites that harvested login credentials. Hackers impersonated VR Systems employees and sent phishing emails containing viruses to election officials in several US states.<sup>60</sup> With AI manipulation, attackers could tailor such emails to specific election officials or deploy AI-generated voice calls exacerbating such attacks.

## F. Mitigation Methods

The expectation that future threats will likely arise as new iterations and permutations of existing risks, instead of completely unknown dangers, suggests that elections can be safeguarded through customization and finetuning of existing processes, strategies, and measures in addition to new methods.<sup>61</sup> This adaptability allows for a proactive approach to AI-enhanced challenges in the election landscape.

### **Proactive Preparedness**

Proactive communication and trust building can help protect against possible AI-enabled election threats. For example, eligible government organizations should host their content on webpages with a .gov domain name to signal trust and safety.<sup>62</sup> Creating solid relationships with media, civil society, and other stakeholders while maintaining an effective crisis communications plan in case of emergency also helps to provide additional assurance against possible threat incidents that may occur. Policy Expert 1 explains, "it is crucial to proactively flood the information space with trusted and authoritative sources, such as election officials' communications. For example, during the European Union (EU) elections, public campaigns directed voters to verified sources like the EU Commission's website."<sup>xi</sup> By proactively flooding social media platforms and the media environment with authoritative and legitimate sources, nations would also be proactively preparing to defend against fake content dissemination.

### **Policy and Government**

Existing laws and regulations should be leveraged to mitigate issues presented by threat actors. For example, in the US the neutrality of laws related to technology and technological developments lends itself to be flexible and adaptable to quickly evolving technologies like AI. Data Consultant 1, a policy and tech translator, product consultant, and long-term digital strategist, suggests that "one of the reasons AI is already being effectively regulated in the US,

---

<sup>xi</sup> Policy Expert 1 interview.

particularly in certain sectors, is due to our existing technology neutral laws. These laws allow various agencies to implement regulations without Congress explicitly granting them authority over AI [...] the financial sector is one major example [and] the health care sector, eligibility, and civil rights: all of those are tech neutral.”<sup>xii</sup> Additionally, as technologies evolve and AI-enabled disruptions increase, pressure from governments and civil society on developers of GenAI technologies will continue to rise as well. Policy approaches and guidance, such as technology neutral laws, are critical to ensuring that developers continue to finetune technologies.

## **Election Security Measures**

Existing approaches to securing elections help mitigate AI-enhanced election risks because AI exacerbates existing threats. The Federal Bureau of Investigation (FBI) oversees pursuing federal election crimes which include campaign finance fraud, civil rights violations, and voter or ballot fraud.<sup>63</sup> Mechanisms such as the Center for Internet Security’s Elections Infrastructure Information Sharing and Analysis Center (EI-ISAC)<sup>64</sup> improve information sharing among election officials. The CISA Government Coordinating Council<sup>65</sup> and the Sector Coordinating Council,<sup>66</sup> created as part of the efforts led by the Department of Homeland Security (DHS) after the 2016 election, share information, identify vulnerabilities, and develop strategies to protect election infrastructure from cyber threats.

Additionally in early September 2024, CISA released election security checklists for both physical security<sup>67</sup> and cybersecurity.<sup>68</sup> The cybersecurity checklist titled “Election Infrastructure Cybersecurity Readiness and Resilience Checklists” specifically mentions that “Election infrastructure and government infrastructure remain attractive targets for a range of malicious actors from cybercriminals to nation state actors.”<sup>69</sup> CISA has also made available Election Security Resources to the public to build American confidence in safe and secure elections.<sup>70</sup>

## **Cybersecurity Tools and Technical Methods**

Time-tested technical security controls are imperative to providing protection, both proactively and reactively, in the face of novel election threats posed by GenAI technology. According to experts, risks including AI-enabled phishing, impersonation and harassment require technical controls, such as Multifactor Authentication (MFA), Domain-based Message Authentication, Reporting and Conformance (DMARC), Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM) and moving toward zero trust security principles.<sup>71</sup> The US is already implementing such methods as the US EAC has implemented a strict configuration of DMARC.<sup>72</sup> On social media platforms, enabling privacy settings, such as

---

<sup>xii</sup> Data Consultant 1 interview.

deleting inactive accounts, is a good cyber hygiene practice that can help protect against emerging threats to elections. The use of human authentication tools—such as CAPTCHAs and physical verifications—is also imperative to differentiating human users from automated processes. Implementing such tools on forms and open records requests, especially website-based submissions, can reduce the volume of inauthentic requests an election office receives.

## **AI-Hardening**

While traditional cybersecurity tools and technical methods can help mitigate many AI-enabled election threats, the constant evolution of AI technologies require such tools and methods to be reviewed and augmented. For example, as AI capabilities are enhanced, authentication tools will need to get sophisticated or “AI-hardened.” AI hardening focuses on solutions specifically designed for AI-related vulnerabilities and includes tactics like safe and secure AI development, AI red teaming, AI auditing, and hardware-linked software actions.<sup>73</sup> An example of AI hardening was seen in the 2024 Taiwanese verification method in which users were verified as humans but remained anonymous in order to combat against AI-generated fake personas.<sup>74</sup>

The constant improvement of GenAI makes developers responsible for developing safe and secure technology. For example, OpenAI utilizes both Rule-Based Rewards (RBRs) and Reinforcement Learning from Human Feedback (RLHF) to ensure the safety and utility of its tools. The company describes RBRs as clearcut rules with simple steps by which AI models can produce output that aligns with safety standards. By plugging RBRs into the traditional RLHF system,<sup>xiii</sup> OpenAI aims to incorporate human inputs while balancing safety and effectiveness as demonstrated in Figure 4 below.<sup>75</sup>

In Figure 4, OpenAI plots the utility and safety of AI technologies, marking the safe and useful region with a shaded area. Two baselines are shown—the human baseline which is very safe and a helpfulness baseline which is useful but less safe. The figure reflects that AI technologies developed with RBR and human reinforcement are safer than those developed only with RBR.

---

<sup>xiii</sup> Traditional RLHF is an AI technique that blends reinforcement learning (RL) with human input to boost learning accuracy and decision-making. Instead of relying solely on trial and error, as in conventional RL where an agent interacts with its environment and is rewarded or penalized based on actions, RLHF integrates expert human feedback. This guidance introduces valuable insights that might not be captured through rewards alone and starts with the agent collecting interaction data, then receiving evaluative feedback from humans. This feedback is incorporated into a reward model, helping the agent interpret actions more effectively. The agent continuously updates its policy using this combined information, iterating the cycle to improve its decision-making and efficiency.

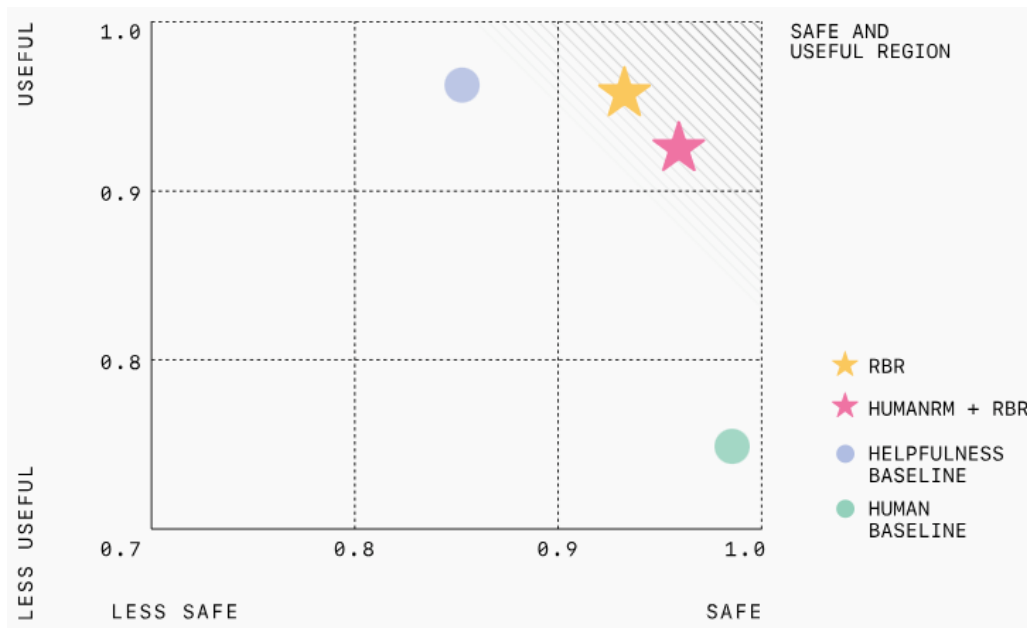


Figure 4: Ensuring Safety and Utility in OpenAI's Technologies<sup>76</sup>

After the development phase, there are various ways to AI-hardened technologies through testing and auditing methods. AI red teaming<sup>77</sup> draws from traditional red teaming practices but may not entail coding or hacking into systems because it focuses on testing for AI-specific vulnerabilities such as jailbreaks and prompt injection attacks. For example, jailbreaking does not require hacking and is the process of stress testing LLMs to force an unethical or harmful output. AI red teaming could encompass adversarial cybersecurity testing of prompt-based applications and underlying models to attempt exfiltrating models and data sets.<sup>78</sup> However, this mitigation currently lacks standardized practices to be able to compare different AI models.

Furthermore, AI auditing is a forthcoming method as well, and NIST is leading the way. NIST launched initiatives to evaluate and audit AI systems. In July 2024, it released the open-source software Dioptra to help users and developers measure how certain attacks can degrade an AI system.<sup>79</sup>

In May 2024, NIST also launched ARIA, a new program to advance sociotechnical testing and evaluation for AI in real world contexts.<sup>80</sup> Reva Schwartz, the lead research scientist,<sup>81</sup> underscores the importance of considering human condition alongside machine performance. ARIA uniquely integrates sociology, psychology, and anthropology into the evaluation process which goes beyond traditional lab-based testing.<sup>82</sup> ARIA will help assess the AI risks with its new set of methodologies and metrics for quantifying how well a system maintains safe functionalities within society contexts. However, the industry and community's focus on ARIA and dedication to using it are not yet clear.



## AI-Enhanced Defenses and Positive Use Cases

Despite the exacerbated election risks due to AI, experts also note that AI can be leveraged to assist in and further secure election processes. These positive use cases include:

- AI to manage voter registration: clean up voter rolls,<sup>83</sup> ensure accurate and up-to-date voter data, efficiently match voter signatures to mail-in ballots<sup>84</sup>
- AI bots to answer questions: respond to misinformation that surfaces on social media,<sup>85</sup> provide real-time information about voting locations and procedures<sup>86</sup>
- AI for voter education: micro-targeted campaigns to help educate voters,<sup>87</sup> translate election materials into various languages, proofread and edit information documents for voters and officials<sup>88</sup>
- AI to prevent cyberattacks: detect abnormalities in election infrastructure<sup>89</sup>

Experts encourage governments using AI to support elections to moderate its usage and ensure “appropriate attention to quality, transparency, and consistency.”<sup>90</sup> The Brennan Center recommends that election officials opt to use the simplest AI technology available: though possibly less sophisticated, simpler tools are easier to use and explain.

The Brennan Center also suggests that bipartisan teams provide human overview of any parts of the election process that use AI technology. For example, if AI is used to match voter signatures to mail-in ballots and a ballot is denied for irregularities or signature mismatches, human review is essential.<sup>91</sup> Officials must also prepare for variability and inconsistencies as these issues may contribute to the proliferation of misinformation, among other concerns. For the time being, adopting AI systems for critical functions in elections is not recommended unless there are already national or state standards in place. As such, it is imperative that effective AI adoption policy is established, appropriate staff training is implemented, and robust contingency plans are put into place.<sup>92</sup>

## G. Summary

Despite social media and news’ depiction of GenAI as creating innumerable new and complex challenges to civil and political society, experts tend to agree that GenAI will not be presenting completely new election threats.<sup>93</sup> The election space was already challenging due to decontextualized videos and images, misinformation, and extreme narratives.

As the Office of the Director of National Intelligence (ODNI) assessed in mid-September 2024, just 45 days before the presidential election, AI is boosting, not revolutionizing, foreign interference in the US election.<sup>94</sup> Notably, the ODNI also reported that Russia has spread the most AI-generated texts, images, audio, and video related to the US election.<sup>95</sup> Regardless,

experts have yet to see widespread concerns about AI-driven deepfakes or deceptive materialize on a large scale.<sup>xiv</sup>

Experts indicate that sociological and cultural issues already dividing civil societies around the world represent bigger issues to the electoral process than GenAI. GenAI simply “amplifies the abilities of all good and bad actors in the system to achieve all the same goals they’ve always had [...] this technology that we’ve developed is going to have effects on our democracy.”<sup>96</sup> Adversaries have utilized AI to amplify or further extremify existing narratives to achieve various objectives.<sup>xv</sup>

Additionally, the risk of overhyping the threat of AI-generated content also exists. There is still limited evidence of and research on the widespread adoption of GenAI tools in the information space. The fear of the undefined threat can create more issues at present than the threats themselves as exemplified by the “Liar’s Dividend” threat<sup>97</sup> as well as the degrading credibility of the information space and electoral processes. In particular with the Liar’s Dividend, governments and industries can implement enhanced content provenance standards, improved deepfake detection technology, strengthened public education and truth discernment, strong norms against false claims, and established public trust in authoritative messengers and information sources.<sup>98</sup>

Despite the AI-enabled election risks, various mitigation methods from proactive preparedness and policy approaches to cybersecurity tolls and technical methods to AI-hardening approaches can help safeguard elections around the world. AI tools can also be utilized to manage voter registration, answer voter questions, provide electoral information, educate voters, and prevent cyberattacks on election infrastructure.

Appendix A at the end of this report includes a table of case studies regarding AI usage in recent elections around the world.

---

<sup>xiv</sup> Policy Expert 1 interview.

<sup>xv</sup> Policy Expert 1 interview.

## 6. AI Governance Measures

### A. Background

GenAI adoption has increased rapidly in the last few years raising concerns about potential risks associated with bias, privacy violations, and unintended consequences. As GenAI continues to transform various sectors, there is a growing consensus among experts, policymakers, and industry leaders that a robust governance structure is essential to ensure its safe and ethical development and deployment.<sup>99</sup>

Governments worldwide are closely scrutinizing the new technology: some have initiated regulations or oversight measures concerning the development of AI tools while others have proceeded cautiously, focused on exploration and research. Various international bodies such as the Organization for Economic Co-operation and Development (OECD) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) have issued guidelines to ensure that AI development follows ethical principles. Governments are also developing their own AI strategies and frameworks, such as the European Union (EU)'s AI Act and Canada's AI and Data Act (AIDA), emphasizing risk management and ethical guidelines. Governments, companies, and nongovernmental organizations (NGOs) are utilizing Public-private collaborations—such as the G7 Hiroshima AI Process—to promote AI governance standards as well.

The debate over AI regulation has intensified, and challenges in striking the right balance exist: Overly prescriptive policies could stifle innovation and slow progress while permissive frameworks might expose society to significant and avoidable risks.<sup>xvi</sup> Effective regulation may include mandates for dataset openness, human oversight, and transparency in commercial GenAI models alongside industry self-regulation and ethical guidelines.

Overall, the AI governance landscape is rapidly evolving and characterized by four broad categories: risk-based, rules-based, principles-based, and outcomes-based, as outlined in Figure 5. However, attributing singular approaches to specific legislation or frameworks is challenging, and hybrid approaches from combining complementing elements from different approaches are created as well. Various expert interviews highlighted that a hybrid approach is the use-case approach which combines risk and outcomes-based approaches.

---

<sup>xvi</sup> AI Policy Expert 2, Data Consultant 1, and Security Expert 1 interview.

	Risk-based	Rules-based	Principles-based	Outcomes-based
<b>Definition</b>	Focuses on classifying and prioritizing risks in relation to the potential harm AI systems could cause	Lays out detailed and specific rules, standards and/or requirements for AI systems	Sets out fundamental principles or guidelines for AI systems, leaving the interpretation and exact details of implementation to organizations	Focuses on achieving measurable AI-related outcomes without defining specific processes or actions that must be followed for compliance
<b>Benefits</b>	<ul style="list-style-type: none"> <li>– Tailored to application area</li> <li>– Proportional to risk profile</li> <li>– Flexible to changing risk levels</li> </ul>	<ul style="list-style-type: none"> <li>– Potential reduction of complexity</li> <li>– Consistent enforcement possible</li> </ul>	<ul style="list-style-type: none"> <li>– Intended to foster innovation</li> <li>– Adaptable to new developments</li> <li>– Can encourage sharing of best practices</li> </ul>	<ul style="list-style-type: none"> <li>– Can support efficiency</li> <li>– Flexible to change</li> <li>– Intended to foster innovation</li> <li>– Compliance can be cost-effective</li> </ul>
<b>Challenges</b>	<ul style="list-style-type: none"> <li>– Risk assessments can be complex</li> <li>– May create barriers to market entry in high-risk areas</li> <li>– Assessment and enforcement can be complex</li> </ul>	<ul style="list-style-type: none"> <li>– Rigidity can increase compliance costs</li> <li>– May be unreliable to enforce</li> </ul>	<ul style="list-style-type: none"> <li>– Potential inconsistencies with interpretation of principles</li> <li>– Unpredictable compliance and impractical enforcement</li> <li>– Potential for abuse by bad actors</li> </ul>	<ul style="list-style-type: none"> <li>– Scope of measurable outcomes can be vague</li> <li>– Potential for diffused accountability</li> <li>– Limited control over process and transparency</li> </ul>
<b>Example</b>	<b>EU:</b> <i>Artificial Intelligence Act, 2023</i> (provisional agreement)	<b>China:</b> <i>Interim Measures for the Management of Generative AI Services, 2023</i>	<b>Canada:</b> <i>Voluntary Code of Conduct for Artificial Intelligence, 2023</i>	<b>Japan:</b> <i>Governance Guidelines for Implementation of AI Principles Ver. 1.1, 2022</i>

Figure 5: Summary of AI governance approaches<sup>100</sup>

Globally, regulators and policymakers are attempting to implement AI governance measures to varying degrees. The EU's recent provisional agreement on the AI Act represents a world-first effort in global AI regulation, focusing on AI products and services through a risk-based and use-case-driven framework.<sup>101</sup> Other nations—such as Canada, Brazil, Chile, and the Philippines—are also developing AI-specific regulations.<sup>102</sup> Meanwhile, India is exploring a non-regulatory approach to foster innovation and adaptation to AI's rapid advancement.<sup>103</sup> Responding to the rise of GenAI models, China has implemented specific regulations while the EU AI Act includes obligations for foundation models supporting general-purpose AI systems. Countries like Singapore, Malaysia, Saudi Arabia, Japan, and Rwanda are formulating national policies to govern AI, employing a spectrum of regulatory instruments from hard laws to voluntary best practices across various sectors.<sup>104</sup> Appendix B provides an overview of current regulations outside of the US.

## B. US AI Regulation and Framework

In the US, comprehensive federal legislation or regulations specifically governing the development or usage of AI in the US do not exist. Despite the absence of comprehensive legislation, various frameworks and guidelines are in place to provide direction on AI governance, and the number of AI-related regulations in the US has been rapidly increasing. For instance, the number of AI-related regulations rose from one in 2016 to 25 in 2023, as shown in Figure 6.

### Number of AI-related regulations in the United States, 2016–23

Source: AI Index, 2024 | Chart: 2024 AI Index report

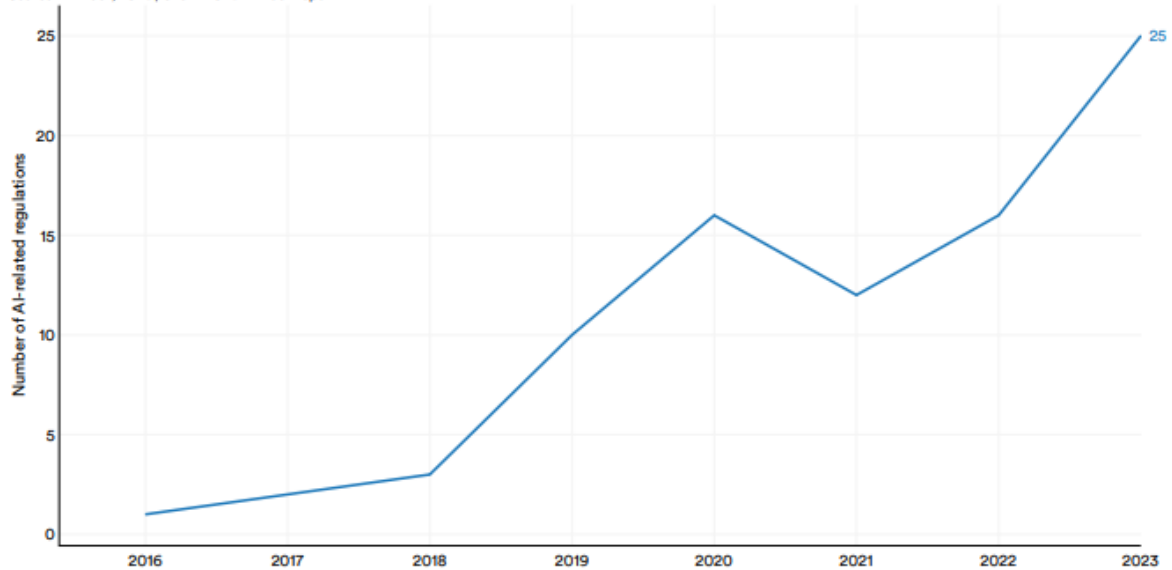


Figure 6: Number of AI-related regulations in the US, 2016-23<sup>105</sup>

Notably, increasing efforts are being made to regulate the use of AI-generated content in elections. Three bills that aim to regulate deepfakes and other AI-generated content in elections have passed the markup stage in the Senate Committee on Rules and Administration and are heading for a final vote.<sup>106</sup> These bills include the Protect Elections from Deceptive AI Act<sup>xvii</sup>, the AI Transparency in Elections Act<sup>xviii</sup>, and the Preparing Election Administrators for AI Act<sup>xix</sup>. The proposed legislation reflects growing concerns about the potential for AI to undermine election integrity and aims to improve transparency and oversight in the use of AI technologies in political campaigns. As of August 2024, the 118th Congress had introduced over 100 AI-related bills. Though none have been enacted yet, there is movement and support from congressmen to formally pass AI related bills.<sup>107</sup> For example, in September 2024, the House Science, Space, and Technology Committee approved nine bipartisan AI-related bills, sending them to review by the House of Representatives.<sup>108</sup> The following section outlines significant approaches in the US to AI regulation at the federal and state level.

#### I. Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of AI

##### Overview

In October 2023, the White House issued Executive Order (EO) 14110—now dubbed the AI

<sup>xvii</sup> The Protect Elections from Deceptive AI Act prohibits the intentional distribution of AI-generated audio or visual media related to federal candidates, especially if intended to influence an election or solicit funds.

<sup>xviii</sup> The AI Transparency in Elections Act establishes labelling standards for political ads, requiring clear disclosure if the ad contains AI-generated images, audio, or video.

<sup>xix</sup> The Preparing Election Administrators for AI Act would require the Election Assistance Commission to team with NIST on a report that delivers voluntary guidelines for election administrators on the related risks and benefits of AI.

EO—on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.<sup>109</sup> This EO outlines comprehensive standards for AI, covering safety, security, privacy, equity, civil rights, workforce development, innovation, competition, and responsible government use. The EO emphasizes cybersecurity, including measures for offensive cyber operations and guidelines for auditing AI capabilities to mitigate potential harm. The EO’s definition of AI systems is broad; it is not limited to GenAI or systems leveraging neural networks but is inclusive of systems that have been built over the last several years.<sup>110</sup>

The AI EO represents a pivotal step by the Biden administration and underscores a dual commitment: promoting a vibrant AI ecosystem that drives economic growth and global competitiveness while also addressing critical governance issues to safeguard against risks and promote trust in AI technologies. The EO directs over 50 federal entities in more than 150 requirements across 13 sections, as detailed in Figure 7 below. Though the EO encompasses a wide range of concerns, specific AI-related risks and policy domains have a greater number of distinct requirements. For instance, three sections—Section 4 (Safety), Section 5 (Innovation), and Section 10 (Government)—comprise approximately two-thirds of all specified requirements. However, raw counts do not convey the comprehensive nature of these requirements as they differ in their scope and complexity.

Full Section Name	Our Shorthand	Summary of Requirements
Sec. 1. Purpose	Purpose	Summary of the EO’s purpose
Sec. 2. Policy and Principles	Principles	Seven overarching principles to guide all federal action outlined in the EO
Sec. 3. Definitions	Definitions	Thirty-three definitions of terms found in the EO (e.g., “agency” and “generative AI”)
Sec. 4. Ensuring the Safety and Security of AI Technology	Safety	Requirements to ensure the safety, security, and reliability of AI systems through standards, evaluations, benchmarking, and information disclosure requirements
Sec. 5. Promoting Innovation and Competition	Innovation	Requirements to promote innovation and competition through immigration reform, investment in resources, support for R&D, and IP protection
Sec. 6. Supporting Workers	Workers	Requirements to increase the government’s understanding of and address AI-related workforce disruptions
Sec. 7. Advancing Equity and Civil Rights	Civil Rights	Requirements to address unlawful discrimination and promote equitable treatment and civil rights in the criminal justice system, government benefits and programs, and the broader economy
Sec. 8. Protecting Consumers, Patients, Passengers, and Students	Consumers	Requirements to protect consumers from AI-related fraud, discrimination, and threats to privacy in healthcare, transportation, education, and telecommunication
Sec. 9. Protecting Privacy	Privacy	Requirements to mitigate privacy risks exacerbated by AI
Sec. 10. Advancing Federal Government Use of AI	Government	Requirements to manage federal government use of AI and plan a national surge in AI talent in the federal government
Sec. 11. Strengthening American Leadership Abroad	International	Requirements to strengthen U.S. leadership of global AI governance efforts including international risk frameworks, technical standards, and multilateral engagement
Sec. 12. Implementation	Implementation	Establishes the White House Artificial Intelligence Council within the Executive Office of the President
Sec. 13. General Provisions	General	Clarifies the EO’s legal effect

Figure 7: Overview of the EO’s sections<sup>111</sup>

Another feature of the EO is the deadlines attached to most requirements. 72% of

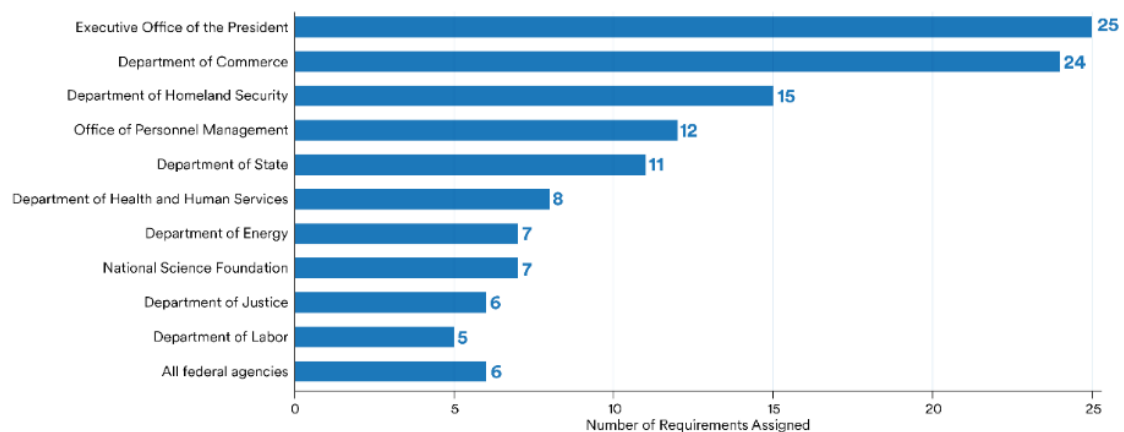
requirements have “time-boxed” deadlines: the actions must be completed by a specified date. Stanford University’s Human-Centered Artificial Intelligence was able to confirm 70-80% of requirements with deadlines between 90 and 180 days had been completed within 180 days.<sup>xx</sup> The rapid turnaround times could indicate that many tasks assigned to federal entities were already in progress. There also may have been an urgency to achieve as much as possible before the end of Biden’s term as a presidential transition might result in the revocation of executive orders.

## Role of Agencies and Departments

The EO mainly assigns tasks to more than 50 federal agencies or entities with prescribed deadlines, and an even larger number of agencies and entities are referenced as supporting functions. The EO also encourages independent federal regulators to engage in given tasks but does not prescribe accompanying deadlines for those tasks. In response to these assignments, many federal entities released their guidance and roadmaps. Figure 8 below details the distribution of requirements across federal entities. The federal entities within the Executive Office of the President (EOP) and the Department of Commerce (DOC) carry the highest number of responsibilities with 25 and 24 requirements, respectively. Notably, the EOP encompasses various agencies each with distinct mandates. The Director of the Office of Personnel Management (OPM) oversees the implementation of 12 requirements, accounting for 48 percent of the EOP's tasks. Other agencies with substantial task lists include the DHS, the Office of Personnel Management (OPM), and the Department of State (DoS). The following section details the various responsibilities of select federal agencies and departments along with recent AI initiatives made to fulfill requirements outlined in the EO.

Distribution of requirements across federal entities\* (Executive Order 14110)

Source: Stanford HAI, RegLab, CRFM, 2023



\* Note: Counts reflect only unambiguously assigned requirements—stakeholders may be required to fulfill additional requirements. We only show entities specifically named as responsible for five or more requirements. Requirements that task a Department Secretary “through” the head of sub-agencies fall within the count for the parent-level Department. We also show requirements assigned to all federal agencies as a separate category.

Figure 8: Distribution of requirements across federal entities<sup>112</sup>

<sup>xx</sup> AI Policy Expert 3 interview.

### Executive Office of the President (EOP)

The EOP was assigned the highest number of EO requirements at 25, detailed in Table 3 below. The EOP is not a single entity but comprises various agencies (e.g., the Council of Economic Advisors), senior officials (e.g., the Assistant to the President for National Security Affairs), permanent councils (e.g., the Chief Data Officers Council), and federal advisory committees (e.g., the President’s Council of Advisors on Science and Technology) with distinct mandates. The Director of the OMB an agency within the EOP, is responsible for overseeing the implementation of 12 tasks, accounting for 48% of the EOP's responsibilities.

Table 3: Overview of EOP EO Requirements and Status of Implementation<sup>113</sup>

Section		Summarized Requirements	Status <sup>114</sup>
Sec. 4.3(a)(iv)	Safety	Coordinate with agency heads to develop mandatory guidelines ensuring the security and resilience of critical infrastructure.	No information available but the deadline is Dec 2024.
Sec. 4.4(b)(i)	Safety	Establish a framework to screen and monitor the procurement of synthetic nucleic acids to prevent misuse.	Implemented.
Sec. 4.5(c)	Safety	Issue detailed guidance to agencies on labeling and authenticating official digital content to enhance security.	In progress: to be after the guidance is developed and likely will not be completed until Jun 2025.
Sec. 4.5(d)	Safety	Consider amendments to the Federal Acquisition Regulation based on digital content labeling guidance.	No information available.
Sec. 4.7(a)	Safety	Develop and disseminate guidelines for security reviews of Federal data that could pose security risks if misused.	No information available but was supposed to be done by Jul 2024.
Sec. 4.8(a)-(b)	Safety	Oversee an interagency process to develop a National Security Memorandum focused on AI and related technologies.	No information available but was supposed to be done by Jul 2024.
Sec. 5.2(h)	Innovation	Submit a comprehensive report on AI's current and future role in advancing scientific research across domains.	Implemented.
Sec. 6(a)(i)	Workers	Submit a detailed report examining the effects of AI on the labor market, including job displacement and creation.	Implemented.



Sec. 9(a)(i)	Privacy	Evaluate how agencies procure and handle commercially available information (CAI), particularly those containing personally identifiable information (PII).	No information available.
Sec. 9(a)(ii)	Privacy	Assess current agency standards for managing CAI with PII to guide the development of potential new guidance.	No information available.
Sec. 9(a)(iii)	Privacy	Issue a Request for Information (RFI) aimed at enhancing privacy impact assessments across agencies.	Implemented.
Sec. 9(a)(iv)	Privacy	Support implementation actions and strategy formulation based on the outcomes of the RFI process.	No information available.
Sec. 10.1(a)	Government	Convene an interagency council to coordinate the use and integration of AI technologies across agencies.	Implemented.
Sec. 10.1(b)	Government	Issue comprehensive guidance on AI use in the Federal Government, including Chief AI Officer roles and risk management practices.	Implemented.
Sec. 10.1(c)	Government	Develop a standardized method for agencies to track, report, and ensure compliance with AI adoption guidelines.	No information available but was supposed to be done by May 2024.
Sec. 10.1(d)(ii)	Government	Ensure that all agency contracts for AI systems are consistent with established AI guidelines and principles.	No information available but was supposed to be done by Sep 2024.
Sec. 10.1(e)	Government	Issue instructions for collecting and reporting agency AI use cases, with annual updates and reviews.	In progress as of Mar 2024.
Sec. 10.2(a)	Government	Identify and prioritize mission-critical areas where AI talent recruitment is most needed within the Federal Government.	Implemented.
Sec. 10.2(b)	Government	Convene an AI and Technology Talent Task Force to monitor AI workforce hiring, retention, and the dissemination of best practices.	Implemented.
Sec. 10.2(b)(i)	Government	Track and regularly report on the Federal Government's AI workforce capacity, identifying gaps and strengths.	Implemented.
Sec. 10.2(b)(ii)	Government	Identify and share best practices for attracting, hiring, retaining, and training AI talent across agencies.	No information available.
Sec. 10.2(b)(iii)	Government	Coordinate AI talent placement through fellowship programs and agency-specific tech talent initiatives.	No information available.
Sec. 10.2(b)(iv)	Government	Convene a regular forum for AI professionals across Federal agencies to collaborate and share knowledge.	No information available.
Sec. 10.2(c)	Government	Implement strategic plans to support the rapid recruitment and deployment of AI talent across the Federal Government.	Implemented.

Department of Commerce (DOC)

The AI EO prescribes the second highest number of requirements to the DOC. Table 4 below outlines the DOC’s 24 requirements and their implementation status.

Table 4: Overview of DOC EO Requirements and Status of Implementation<sup>115</sup>

Section		Summarized Requirements	Status <sup>116</sup>
Sec. 4.1(a)(i)	Safety	Develop guidelines and best practices for safe and trustworthy AI such as resources for GenAI, secure development practices, and AI benchmarks.	In progress: started in Nov 2023.
Sec. 4.1(a)(ii)	Safety	Establish guidelines for AI red-teaming, including coordination for safety assessments and development of testing environments.	In progress: was supposed to be done by Jul 2024.
Sec. 4.2(a)(i)	Safety	Require companies to report on dual-use foundation models, including development activities, cybersecurity measures, and red-team test results.	Implemented.
Sec. 4.2(a)(ii)	Safety	Require reporting of large-scale computing clusters, their acquisition, development, and computing power.	Implemented.
Sec. 4.2(b)	Safety	Define and update technical conditions for reporting models and computing clusters, including thresholds for computational power and data use.	No information available.
Sec. 4.2(c)	Safety	Propose regulations for Infrastructure as a Service (IaaS) Providers to report foreign transactions involving large AI models and require reporting from foreign resellers.	Implemented.
Sec. 4.2(d)	Safety	Propose regulations for verifying the identity of foreign persons obtaining IaaS accounts, including documentation and access controls.	Implemented.
Sec. 4.4(b)(ii)	Safety	Engage with stakeholders to develop specifications and best practices for nucleic acid synthesis screening.	Implemented.
Sec. 4.5(a)	Safety	Submit a report on standards, tools, and practices for authenticating and labeling synthetic content, and preventing harmful AI-generated content.	In progress: was supposed to be done by June 2024; NIST AI 100-4 issued.
Sec. 4.5(b)	Safety	Develop guidance on digital content authentication and synthetic content detection measures.	No information available but is supposed to be done by Dec 2024.
Sec. 4.6(a)	Safety	Solicit input on risks and benefits of widely available dual-use foundation models and recommend policies for managing them.	No information available but was supposed to be done by Jul 2024.
Sec. 4.6(b)	Safety	Submit a report on the implications of widely available dual-use foundation models and propose regulatory recommendations.	No information available but was supposed to be done by Jul 2024.
Sec. 5.2(c)(i)	Innovation	Publish guidance on AI inventorship and the use of AI in the inventive process, including examples for USPTO patent examiners.	Implemented.

Sec. 5.2(c)(ii)	Innovation	Issue additional guidance on AI and IP considerations, including patent eligibility updates.	In progress: was supposed to be done by July 2024; USPTO released a RFI in Apr 2024.
Sec. 5.2(c)(iii)	Innovation	Consult with the Copyright Office and recommend actions on copyright issues related to AI.	No information available.
Sec. 5.3(b)(i)	Innovation	Implement a flexible membership structure for the National Semiconductor Technology Center to include diverse industry participants.	No information available.
Sec. 5.3(b)(ii)	Innovation	Promote mentorship programs to increase participation in the semiconductor industry, including from underserved communities.	No information available.
Sec. 5.3(b)(iii)-(iv)	Innovation	Increase resources for startups, including funding, datasets, and technical assistance, and consider competition measures in funding notices.	No information available.
Sec. 9(b)	Privacy	Create guidelines for evaluating differential privacy protections, particularly for AI systems.	In progress: NIST issued draft of SP 800-226, should be completed by Oct 2024.
Sec. 10.1(d)(i)	Government	Develop guidelines, tools, and practices for implementing risk-management practices.	No information available but was supposed to be done by Jun 2024.
Sec. 11(b)	International	Lead global efforts to develop AI-related standards, including terminology, best practices, and risk management.	In progress: NIST released a draft plan for Global Engagement in Apr 2024.
Sec. 11(b)(i)	International	Establish a global engagement plan for AI standards development, covering nomenclature, data handling, and risk management.	In progress: on May 2024, the Secretary of Commerce shared plans to launch a global scientific network for AI safety at the AI Seoul Summit.
Sec. 11(b)(ii)	International	Report to the President on actions taken to advance global AI standards.	No information available but should be done by Jan 2025.
Sec. 11(b)(iii)	International	Ensure global AI standards efforts align with the NIST AI Risk Management Framework and national standards strategy.	In progress: various NIST drafts released.

Within the DOC, the NIST, the Bureau of Industry and Security (BIS), the National Telecommunications and Information Administration (NTIA), and the US Patent and Trademark Office (USPTO) are tasked with implementing a substantial portion of the objectives outlined in the EO. NIST released four draft publications aimed at enhancing the safety, security, and reliability of AI systems (see section 5.B.III for more information on NIST AI publications). NIST also piloted NIST GenAI, a challenge series designed to assess GenAI developed by the global research community, in April 2024<sup>117</sup> The series aims to support the development of methods to distinguish human-produced and AI-generated content.<sup>118</sup>

In addition, the USPTO has invited public comments to understand how AI affects their decisions about whether an invention is patentable under US law.<sup>119</sup> This includes questions about what constitutes prior art and how to evaluate the skill level of an ordinary person in the relevant field. Based on these responses, USPTO expects the response to help them evaluate the need for further guidance on these matters, aid the development of guidance and inform USPTO's work in the courts, and provide technical advice to Congress.<sup>120</sup>

The EO also directs the DOC to draft regulations mandating comprehensive reporting requirements concerning AI, including US Infrastructure-as-a-Service (IaaS) providers involved in transactions with foreign persons for training a large AI model (Sec. 4.2(c)).<sup>121</sup> This requirement would implement the directives of the AI EO as well as EO 13984,<sup>xxi</sup> mandating US IaaS providers to identify information about foreign customers through a Customer Identification Program (CIP) and report the implementation of the CIP to the DOC.

Additionally, US IaaS providers must ensure that their foreign resellers comply with CIP and reporting obligations. The proposed rule also mandates reporting on transactions involving the training of large AI models that could facilitate malicious cyber-enabled activities. Moreover, it grants the Secretary authority to potentially prohibit US IaaS product transactions with foreign individuals or jurisdictions.

---

<sup>xxi</sup> EO13984 "Taking Additional Steps to Address the National Emergency with Respect to Significant Malicious Cyber Enabled Activities" was signed in January 2021 aimed at addressing the use of US IaaS products by foreign malicious cyber actors.

## Department of Homeland Security (DHS)

The DHS was tasked as the responsible stakeholder on 15 requirements in AI EO and a supporting stakeholder on another 11 requirements. Table 5 below outlines the requirements for which DHS is a leading stakeholder as well as the implementation status of these requirements.

Table 5: Overview of DHS EO Requirements and Status of Implementation<sup>122</sup>

Section		Summarized Requirements	Status <sup>123</sup>
Sec. 4.3(a)(iii)	Safety	Integrate AI Risk Management Framework (RMF) and security guidance into safety and security protocols for critical infrastructure.	Implemented.
Sec. 4.3(a)(v)	Safety	Establish an AI Safety and Security Board to advise on AI-related security and resilience for critical infrastructure.	Implemented.
Sec. 4.3(b)(ii)	Safety	Develop and complete pilot projects to test and deploy AI capabilities for identifying and fixing vulnerabilities in government systems.	Implemented.
Sec. 4.3(b)(iii)	Safety	Report on pilot projects, including vulnerabilities found and fixed, and lessons learned for effective AI deployment in cyber defense.	No information available but should have been done by Jul 2024.
Sec. 4.4(a)(i)	Safety	Evaluate additional AI risks related to chemical, biological, radiological, and nuclear (CBRN) threats, consult with experts and report on AI models presenting CBRN risks and regulatory recommendations. <sup>xxii</sup>	Implemented
Sec. 4.4(b)(iv)	Safety	Develop and report on a framework for evaluating nucleic acid synthesis screening, including recommendations for strengthening procurement screening.	No information available but should be done by Oct 2024.
Sec. 5.1(a)	Innovation	Streamline visa processes and ensure availability for noncitizens in AI and critical technologies.	Implemented
Sec. 5.1(d)(i)	Innovation	Review and update immigration policies for experts in AI and emerging technologies.	No verifiable implementation information available but should have been done by April 2024.
Sec. 5.1(d)(ii)	Innovation	Continue modernizing the H-1B program and consider rulemaking for adjusting status to permanent residency for tech experts and their families.	Implemented

<sup>xxii</sup> Within DHS, the Countering Weapons of Mass Destruction Office (CWMD) leads the DHS efforts against CBRN threats. The AI EO also specifies that the DHS should collaborate with the OSTP for the EvaluationAI misuse for CBRN threat production.

Sec. 5.1(f)	Innovation	Use discretionary authorities to support and attract foreign nationals with skills in AI.	Implemented
Sec. 5.1(g)	Innovation	Develop and publish resources to attract and retain AI experts, including a comprehensive guide and a public report on immigration system usage.	In progress: the White House and DHS claim to be done, but external sources suggest that it is still in progress.
Sec. 5.2(d)	Innovation	Create a program to combat IP risks related to AI technologies, including personnel for IP theft investigation, coordination with law enforcement, and developing guidance for the private sector.	Implemented.
Sec. 11(d)	International	Lead international efforts to prevent and respond to AI-related disruptions in critical infrastructure.	No information available.
Sec. 11(d)(i)	International	Develop a multilateral plan to promote AI safety and security guidelines for critical infrastructure. <sup>xxiii</sup>	No information available but should have been done by Jul 2024.
Sec. 11(d)(ii)	International	Report on actions to mitigate cross-border risks to U.S. critical infrastructure.	No Information Available but should be done by Jan 2025.

The DHS took various approaches in responding to the AI EO requirements. First, the DHS introduced sector-specific initiatives such as guidelines to mitigate AI risks to critical infrastructure and a report focusing on AI misuse in the development and production of CBRN materials.<sup>124</sup> The DHS also announced initiatives focused on AI security and safety:<sup>125</sup>

- **AI Safety and Security Advisory Board (AISSB):** composed of private and public sector AI experts advising on resilience and incident response for AI in critical infrastructure.
- **Protecting Critical Infrastructure and Cybersecurity:** DHS integrates AI safety guidelines, adapting national frameworks to mitigate risks from AI-enhanced attacks and system failures in critical infrastructure. As mentioned above, DHS established the AISSB with over 20 technology and critical infrastructure executives, civil rights leaders, academics and policy makers to advise on safe and secure development and deployment of AI in critical infrastructure. The DHS also collaborated with CISA to develop the First AI Guidelines for Critical Infrastructure Owners and Operators.<sup>126</sup>
- **Researching Adversarial AI Use:** DHS explores defenses against AI-based threats including biological and chemical risks, collaborating with Countering Weapons of Mass Destruction (CWMD) on risk assessments and mitigation plans.
- **Combatting AI-related Intellectual Property Theft:** DHS develops programs and

<sup>xxiii</sup> For example, at the [AI Seoul Summit 2024](#), South Korea, Japan, Singapore, US, UK, and other Asia-Pacific nations highlighted the need for international cooperation on AI safety standards, especially to protect AI-reliant critical infrastructure. Additionally in April 2024, the US and UK signed a [Memorandum of Understanding \(MOU\)](#) to partner on AI safety, development, and testing although the MOU did not specify critical infrastructure.

guidance to protect AI-related intellectual property, updating enforcement strategies to address emerging threats.

- **Attracting and Retaining Talent in AI and Emerging Technologies:** enhances immigration pathways and processing times for noncitizens contributing to AI and emerging technologies.

Finally, the DHS has also established several principles and initiatives that promote the responsible use of AI:

- **Policy Statement 139-06:** Acquisition and Use of AI and ML by DHS Components: establishes foundational principles for AI use at DHS, ensuring compliance with constitutional and legal standards. Prohibits AI systems from making biased decisions based on protected characteristics.
- **Policy Statement 139-07:** Use of Commercial GenAI Tools: provides guidelines for DHS employees using commercial GenAI tools, emphasizing data safeguarding, privacy protection, and mandatory training on responsible AI use.
- **Directive 026-11:** Use of Face Recognition and Face Capture Technologies: mandates rigorous testing of face recognition technologies to prevent unintended biases or impacts, with ongoing evaluation to meet performance standards.<sup>xxiv</sup>

#### Cybersecurity and Infrastructure Security Agency (CISA)

Within the DHS, CISA holds several key roles of the EO. CISA, along with agency heads with regulatory authority over critical infrastructure and relevant Sector Risk Management Agencies (SRMAs), was tasked with providing an assessment to the Secretary of Homeland Security. The assessment focuses on potential risks associated with AI in critical infrastructure sectors, including how AI deployment might increase vulnerabilities to system failures, physical attacks, and cyber threats. CISA has completed a pilot for this AI-Enabled Vulnerability Detection in July 2024.<sup>127</sup> CISA has also developed a “Roadmap for AI”, a comprehensive framework guiding its AI initiatives. This roadmap not only aligns with the AI EO’s whole-of-government approach and key actions but also includes additional efforts to enhance AI security and support critical infrastructure stakeholders in adopting AI technologies.<sup>128</sup> CISA’s roadmap outlines five lines of efforts (LOEs)<sup>129</sup>:

- **LOE 1: Responsibly use AI to support mission:** CISA will employ AI-enabled software

---

<sup>xxiv</sup> While there has yet to be clear public statements, US AISI may adopt stricter evaluation protocols for AI systems, particularly facial recognition, to align with the DHS mandate for testing to prevent unintended biases and impacts. AISI is likely to focus on developing more detailed cybersecurity standards for facial recognition and face capture technologies reflecting DHS’s emphasis on protecting these systems against cybersecurity threats. This might include guidance on secure data storage, encryption, and risk management for AI systems handling sensitive biometric information.

tools to bolster cyber defense and advance critical infrastructure objectives. The adoption of AI will prioritize ethical, secure, and lawful usage by constitutional mandates and relevant federal policies.

- **LOE 2: Assure AI systems:** CISA will evaluate and support the adoption of secure-by-design AI software across diverse stakeholders, including federal civilian agencies, private sector entities, and state, local, tribal, and territorial governments. This effort includes developing best practices and guidance for resilient AI software development and implementation. CISA will also incorporate the NIST AI Risk Management Framework 1.0 under this LOE.
- **LOE 3: Protect Critical Infrastructure from malicious use of AI:** CISA will assess and recommend strategies to mitigate AI-related threats to the nation's critical infrastructure. Collaboration with government agencies and industry partners involved in AI tool development, testing, and evaluation will be crucial in this LOE.
- **LOE 4: Collaborate and communicate on key AI efforts with the interagency, international partners, and the public:** CISA will engage in DHS-led and interagency initiatives concerning AI-enabled software, contributing to policy development for the US national strategy on AI. CISA will also coordinate with international partners to advance global AI security standards.
- **LOE 5: Expand AI expertise in the workforce:** CISA will enhance AI knowledge within its workforce by providing education on AI. Efforts will include recruiting interns, fellows, and employees with AI expertise, ensuring comprehensive training across legal, ethical, policy, and technical aspects of AI-based software systems.

In alignment with its AI Roadmap, CISA conducted a tabletop exercise with the Joint Cyber Defense Collaborative (JCDC) in June 2024. This exercise supported the development of an AI Security Incident Collaboration Playbook spearheaded by the JCDC.AI—the organization focusing on building a community of AI providers, security vendors, and critical infrastructure operators to address AI-related risks and threats. The exercise involved more than 50 AI experts from across the public and private sectors.<sup>130</sup> CISA will incorporate the lessons learned from this exercise into an AI Security Incident Collaboration Playbook to inform operational collaboration across government, industry, and international partners. A subsequent tabletop exercise will test and validate the Playbook with AI companies and critical infrastructure entities that are integrating AI in their operational environments.



### Office of Personnel Management (OPM)

The OPM was assigned 12 requirements in the EO, all of which are within Section 10: Advancing Federal Government Use of AI which includes requirements to manage federal government use of AI and plan to increase AI talent in federal government. Most of these requirements have already been implemented.

Table 6: Overview of OPM EO Requirements and Status of Implementation<sup>131</sup>

Section		Requirements (Summary)	Status <sup>132</sup>
Sec. 10.1(f)(iii)	Government	Create guidance on GenAI use for federal workforce.	Implemented.
Sec. 10.2(d)(i)	Government	Conduct a review on hiring and workplace flexibility for AI-related roles and authorize direct-hire authority if necessary.	Implemented.
Sec. 10.2(d)(ii)	Government	Consider authorizing temporary excepted service appointments to meet staffing needs for implementing AI-related directives.	Implemented.
Sec. 10.2(d)(iii)	Government	Coordinate a pooled hiring initiative using skills-based assessments to recruit AI talent across various federal agencies.	Implemented.
Sec. 10.2(d)(iv)	Government	Issue guidance on using pay flexibilities and incentive programs to attract and retain AI and other key technical talent.	Implemented.
Sec. 10.2(d)(v)	Government	Establish guidance for skills-based hiring practices to increase access to AI and technology roles for candidates with nontraditional academic backgrounds.	Implemented.
Sec. 10.2(d)(vi)	Government	Form an interagency working group to support government-wide hiring of individuals with AI and other technical skills.	Implemented.
Sec. 10.2(d)(vii)	Government	Review and update Executive Core Qualifications (ECQs) for Senior Executive Service (SES) positions to include AI literacy and related competencies and implement these new ECQs.	Implemented.
Sec. 10.2(d)(viii)	Government	Review AI-related competencies for civil engineers and similar occupations to ensure the Federal Government reflects the increased use of AI in critical infrastructure.	Implemented.
Sec. 10.2(d)(ix)	Government	Collaborate with the Security, Suitability, and Credentialing Performance Accountability Council to assess and streamline personnel-vetting processes for AI and other emerging technologies.	No information available.

Department of State (DOS)

The DOS was listed as the responsible stakeholder for 11 of the EO’S requirements and a supporting stakeholder on another 16.

Table 7: Overview of DOS EO Requirements and Status of Implementation<sup>133</sup>

Section		Requirements (Summary)	Status <sup>134</sup>
Sec. 5.1(a)	Innovation	Streamline visa processes for noncitizens in AI and emerging technologies, ensuring timely processing and appointment availability.	Implemented
Sec. 5.1(b)(i)	Innovation	Consider new criteria for designating countries and skills on the Exchange Visitor Skills List for J-1 nonimmigrants.	White House claims that implementation is complete, but it cannot be verified.
Sec. 5.1(b)(ii)	Innovation	Consider updating the 2009 Revised Exchange Visitor Skills List.	White House claims that implementation is complete, but it cannot be verified.
Sec. 5.1(b)(iii)	Innovation	Consider implementing a domestic visa renewal program for qualified applicants to avoid work interruptions.	Implemented.
Sec. 5.1(c)(i)	Innovation	Consider expanding domestic visa renewal program categories to include J-1 research scholars and STEM F-1 students.	No information available but should have been completed by Apr 2024.
Sec. 5.1(c)(ii)	Innovation	Establish a program to attract top global talent in AI and other technologies to the US and inform them about visa options and expedited adjudication.	No information available but should have been completed by Apr 2024.
Sec. 5.1(f)	Innovation	Use discretionary authorities to attract foreign nationals with skills in AI and other critical technologies.	Implemented.
Sec. 11(a)(i)	International	Lead global engagement to expand understanding of US AI policies and enhance international collaboration.	No information available.
Sec. 11(a)(ii)	International	Develop an international framework for AI risk management and encourage support for voluntary commitments from allies.	No information available.
Sec. 11(c)(i)	International	Publish an AI in Global Development Playbook incorporating AI Risk Management Framework principles for international contexts.	In progress: in Jan 2024, USAID issued an RFI, supposed to be done by Oct 2024.
Sec. 11(c)(ii)	International	Develop a Global AI Research Agenda with guidelines for responsible AI development and recommendations on labor-market implications.	In progress: in Jan 2024, USAID issued an RFI, supposed to be done by Oct 2024.

Beyond the initiatives and tasks in the chart above, in July 2024, the DOS released the "Risk

Management Profile for Artificial Intelligence and Human Rights." The Profile aims to guide organizations—including governments, businesses, and civil society—in aligning AI practices with international human rights standards and in integrating human rights considerations into AI risk management practices, addressing unintentional human rights violations (e.g., biased AI outputs) and intentional abuses (e.g., mass surveillance).<sup>135</sup>

Anchored in international human rights standards, the Profile seeks to provide a unified tool for stakeholders globally to enhance their AI risk management while safeguarding human rights. The Profile incorporates actions from the NIST AI RMF showing how these actions can support human rights due diligence.<sup>136</sup> Additionally, the Profile aligns human rights actions with the AI RMF's organizational functions, ensuring that human rights considerations are embedded throughout the AI lifecycle and across various applications and sectors.<sup>137</sup>

## Department of Health and Human Services (HHS)

The EO tasked the HHS with eight requirements, listed in Table 8. Directives for HHS emphasize its role in multi-agency efforts to enhance national security and support AI research in health care. Key concerns include biosecurity risks in AI-driven synthetic genetic material development and the need for AI evaluation tools to safeguard data and models. HHS is crucial in developing the National AI Research Resource (NAIRR), a pilot program offering computational resources, data, and support to AI researchers.<sup>138</sup> The EO gives HHS authority to ensure the safe deployment of AI in healthcare, enforce compliance with federal nondiscrimination laws, oversee AI use in drug development via the Food and Drug Administration (FDA), and support responsible AI development through National Institute of Health (NIH) research and private sector collaboration.

Table 8: Overview of HHS EO Requirements and Status of Implementation<sup>139</sup>

Section		Requirements (Summary)	Status <sup>140</sup>
Sec. 5.2(e)	Innovation	Advance responsible AI innovation in healthcare by identifying and prioritizing grantmaking and awards to support responsible AI development, focusing on personalized immune-response tools, improving healthcare-data quality, and accelerating grants for health equity in underserved communities.	No information available.
Sec. 7.2(b)(i)	Civil Rights	Promote equitable administration of public benefits with a plan addressing automated systems in public benefits programs, ensuring access, human oversight, and fairness.	Implemented.
Sec. 8(b)(i)	Consumers	Establish an HHS AI Task Force and develop a strategic plan for responsible AI deployment in health and human services, focusing on healthcare delivery, public health, equity, safety, and AI-enhanced cybersecurity.	Implemented.
Sec. 8(b)(i)	Consumers	Develop policies and frameworks for responsible AI deployment in healthcare, including safety, equity, privacy, and collaboration with local agencies to advance AI best practices.	White House claims it was done in Apr 2024, but it cannot be verified.
Sec. 8(b)(ii)	Consumers	Direct HHS components to develop a strategy to ensure AI technologies maintain quality in healthcare, including the development of AI assurance policies and infrastructure for pre- and post-market oversight.	White House claims it was done in Apr 2024, but it cannot be verified.
Sec. 8(b)(iii)	Consumers	Promote understanding and compliance with Federal nondiscrimination laws by health providers using AI, through technical assistance, guidance, or other actions as necessary.	Implemented.
Sec. 8(b)(iv)	Consumers	Establish an AI safety program with Patient Safety Organizations to identify, track, and analyze clinical errors from AI in healthcare as well as disseminate best practices to avoid bias and discrimination. <sup>xxv</sup>	No information available.
Sec. 8(b)(v)	Consumers	Develop a strategy for regulating AI tools in drug development, defining objectives, identifying areas for rulemaking, and assessing resources to implement a regulatory system while considering other identified risks.	No information available.

<sup>xxv</sup> While the HHS has not explicitly mentioned incorporating ISO standards for patient safety, the HHS has developed a trust and safety playbook on AI: [Trustworthy AI \(TAI\)](#). The HHS also has [“Example HHS Use Cases”](#) to satisfy the EO requirement on creating an inventory of non-classified and non-sensitive current and planned AI use cases.

## Department of Energy (DOE)

The EO outlines seven requirements for the DOE, listed in Table 9. The DOE will lead several initiatives including developing AI risk mitigation tools for nuclear security and critical infrastructure, collaborating with other agencies and sectors to build AI models, training 500 new AI researchers by 2025 with NSF's help, and establishing an office to coordinate AI efforts across its programs and National Laboratories.

Table 9: Overview of DOE EO Requirements and Status of Implementation<sup>141</sup>

Section		Requirements (Summary)	Status <sup>142</sup>
Sec. 4.1(b)	Safety	Develop AI model evaluation tools and testbeds at the DOE to assess and mitigate security threats related to AI.	No information available but should have been done by Jul 2024.
Sec. 5.2(b)	Innovation	Establish a pilot program to train 500 AI researchers by 2025, enhancing high-performance and data-intensive computing.	Implemented.
Sec. 5.2(g)(i)	Innovation	Issue a public report on AI's potential to improve electric grid infrastructure planning and operations, and support a clean, resilient energy economy.	Implemented.
Sec. 5.2(g)(ii)	Innovation	Develop tools to build foundation models for science, streamlining permitting and environmental reviews while improving outcomes.	Implemented.
Sec. 5.2(g)(iii)	Innovation	Collaborate with private sector and academia to develop AI tools that mitigate climate change risks.	In progress: White House claimed it was done by Apr 2024, but DOE announced otherwise in Apr 2024.
Sec. 5.2(g)(iv)	Innovation	Expand partnerships to utilize DOE's computing capabilities and AI testbeds for new applications in science, energy, and national security, including climate resilience and clean-energy deployment. Expand partnerships to utilize DOE's computing capabilities and AI testbeds for new applications in science, energy, and national security, including climate resilience and clean-energy deployment.	Implemented.
Sec. 5.2(g)(v)	Innovation	Establish an office to coordinate AI and other critical technology development across DOE programs and National Laboratories.	Implemented.

## National Science Foundation (NSF)

The NSF was given eight requirements by the EO, all of which are summarized in Table 10. These requirements focus on advancing AI research, fostering innovation, and ensuring ethical practices in AI deployment.

Table 10: Overview of NSF EO Requirements and Status of Implementation<sup>143</sup>

Section		Requirements (Summary)	Status <sup>144</sup>
Sec. 5.2(a)(i)	Innovation	Launch a pilot program for the National AI Research Resource (NAIRR) to support AI-related research and development.	Implemented.
Sec. 5.2(a)(ii)	Innovation	Fund and launch an NSF Regional Innovation Engine that prioritizes AI-related work.	Implemented.
Sec. 5.2(a)(iii)	Innovation	Establish four new National AI Research Institutes.	No information available but should be done by Apr 2025.
Sec. 6(c)	Workers	Prioritize resources for AI-related education and workforce development through existing programs.	In progress: NSF launched the EducateAI Initiative in Dec 2023.
Sec. 9(c)(i)	Privacy	Fund and create the Research Coordination Network (RCN) to advance privacy research and Privacy Enhancing Technologies (PETs) development.	Implemented.
Sec. 9(c)(ii)	Privacy	Engage with agencies to identify and incorporate PETs into their operations, prioritizing research for PETs adoption.	No information available.
Sec. 9(c)(iii)	Privacy	Use the US-UK PETs Prize Challenge results to inform PETs research and adoption approaches.	No information available.

Department of Justice (DOJ)

The DOJ is assigned six requirements, all of which are in Section 7, Advancing Equity and Civil Rights.

Table 11: Overview of DOJ EO Requirements and Status of Implementation<sup>145</sup>

Section		Requirements (Summary)	Status <sup>146</sup>
Sec. 7.1(a)(i)	Civil Rights	Coordinate with agencies to enforce federal laws addressing civil rights, civil liberties violations, and discrimination related to AI.	No information available.
Sec. 7.1(a)(ii)	Civil Rights	Convene a meeting of federal civil rights offices to address AI-related discrimination, increase coordination, and improve public awareness.	Implemented.
Sec. 7.1(a)(iii)	Civil Rights	Provide guidance and training to State, local, Tribal, and territorial investigators on best practices for addressing AI-related civil rights violations.	No information available.
Sec. 7.1(b)	Civil Rights	Report to the President on AI in the criminal justice system, identifying areas for improvement and recommending best practices for law enforcement, including safeguards for AI use.	No information available but should be done by Oct 2024.
Sec. 7.1(c)(ii)	Civil Rights	Develop recommendations for law enforcement on recruiting/training staff with AI knowledge, consulting with state, local, tribal, and territorial agencies.	No information available but should have been done by Jul 2024.
Sec. 7.1(c)(iii)	Civil Rights	Review and reassess the capacity to investigate AI-related civil rights violations by law enforcement, including through improved training for federal officers and prosecutors.	No information available but should be done by Oct 2024.

## Department of Labor (DOL)

The DOL is the lead for five requirements outlined in Table 12. Most of the DOL's AI-related initiatives focus on providing employers with guidelines for implementing AI technology with a strong emphasis on enhancing job quality and safeguarding workers' rights.

Table 12: Overview of DOL EO Requirements and Status of Implementation<sup>147</sup>

Section		Requirements (Summary)	Status <sup>148</sup>
Sec. 5.1(e)	Innovation	Publish a request for information (RFI) to gather input on AI and STEM-related occupations lacking sufficient qualified US workers, for potential updates to the "Schedule A" list.	Implemented.
Sec. 6(a)(ii)	Workers	Submit a report analyzing federal programs' ability to support workers displaced by AI and suggest measures to strengthen or develop support.	White House claims it was done in Apr 2024, but it cannot be verified.
Sec. 6(b)(i)	Workers	Develop and publish best practices for employers to mitigate AI's potential harms and maximize its benefits for employee well-being, including job displacement and labor standards.	Implemented.
Sec. 6(b)(iii)	Workers	Issue guidance for employers using AI to monitor or augment work comply with the Fair Labor Standards Act and legal worker compensation protections.	Implemented.
Sec. 7.3(a)	Civil Rights	Publish guidance for federal contractors on nondiscrimination in AI-based hiring systems to prevent unlawful discrimination.	Implemented.



### Department of Defense (DOD)

According to the AI EO, the DOD is tasked as the leading agency for four requirements and the supporting entity for another nine.

Table 13: Overview of DOD EO Requirements and Status of Implementation<sup>149</sup>

Section		Requirements (Summary)	Status <sup>150</sup>
Sec. 4.3(b)(ii)	Safety	Complete operational pilot projects to remediate AI-related vulnerabilities in federal systems.	In progress: DHS piloted programs in Apr 2024.
Sec. 4.3(b)(iii)	Safety	Report results of AI pilot projects, including vulnerabilities found and fixed, to the Assistant to the President for National Security Affairs.	No information but should have been done by Jul 2024.
Sec. 4.4(a)(ii)	Safety	Conduct a study on AI's impact on biosecurity, including risks from GenAI and dataset use, and make recommendations for mitigation.	In progress: White House claims it was done in Mar 2024, but the National Academies of Science Engineering and Medicine claims it is progress as of April 2024.
Sec. 10.2(h)	Government	Report to the President with recommendations on improving the recruitment and retention of noncitizens with AI expertise, including streamlining access to classified information and enlistment processes.	White House claims it was done in Apr 2024, but it could not be verified.

In addition, the AI EO invokes the Defense Production Act (DPA) which grants the President sweeping authorities to compel or incentivize industry in the interests of national security.<sup>151</sup> For example, Section 4.2 of the EO invokes the DPA's Title VII authorities, which allows the government to compel companies to provide information to the government.<sup>152</sup> It delegates the Secretary of Commerce the authority to require companies that are developing, or showing an intention to develop, potential dual-use foundation models to submit specific information to the government, including information from red-teaming.

### Department of Treasury

The AI EO assigns the Department of Treasury as the lead agency for one requirement: Section 4.3(a)(iii). The Treasury is to "issue a public report on best practices for financial institutions to manage AI-specific cybersecurity risks."<sup>153</sup> In March 2024, Treasury released a report on Managing AI-Specific Cybersecurity Risks in the Financial Services Sector.<sup>154</sup> The report examines the current landscape of AI-related cybersecurity and fraud risks within the financial services sector. The report includes an overview of existing financial AI use cases, trends in threats, and risks. Notably, there is concern that AI tools used to identify fraud are not reliable yet as well as concerns that cybercriminals are using AI to impersonate victims and conduct fraud. Experts have assessed that the report is one of the most important and specific AI governance documents as it covers a broad range of topics and provides an example for state regulators to consider in state-level AI governance measures.<sup>xxvi</sup>

<sup>xxvi</sup> Government Agency 1 and Election Analyst 1 interviews.

## Overlaps and Synergies Among Agencies & Departments

The EO emphasizes a “whole of government” approach that aims to position the US as a leader in AI development and deployment while aligning with international standards for ethical AI practices.<sup>155</sup> As such, the EO takes a collaborative approach, assigning lead and supporting roles for agencies and departments across many requirements. Of the 150 requirements, 77 have one or more agencies or departments assigned to coordinate with or consult with the leading stakeholder.

Across the EO, certain policy issue areas reveal distinct patterns of collaboration, as shown in Table 14 below. Section 4 – Safety stands out as the issue with the most collaborations or coordination efforts, followed by Section 10 – Government, and Section 11 – International. Some examples of overlapping responsibilities include the DOC and DOS frequently tasked together with requirements within the Safety section. Likewise, DHS and the DOS often coordinate on the Innovation and International sections and are tasked with collaborating on global leadership and immigration policy related to AI. The most frequent collaborations between agencies are listed in Table 14 below.

Table 14: Key Agency Collaborations in the EO

Collaboration	Description of interaction
OPM and OMB (within the Executive Office)	<ul style="list-style-type: none"> <li>Review and enhance hiring flexibility and recruitment strategies for AI talent.</li> <li>Streamline and strengthen vetting processes for AI and emerging technology.</li> </ul>
DOC and DOS	<ul style="list-style-type: none"> <li>Coordinate efforts to evaluate dual-use AI models and assess associated risks.</li> <li>Consult with the US Copyright Office to for AI-related copyright issues.</li> <li>Advance global AI standards through international collaboration.</li> </ul>
DHS and DOS	<ul style="list-style-type: none"> <li>Coordinate efforts to update immigration policies to attract AI experts.</li> <li>Lead international initiatives to prevent, respond to, and recover from AI-related risks to critical infrastructure.</li> <li>Develop a multilateral engagement plan to globally promote AI safety and security guidelines.</li> </ul>
DOC (NIST) and DHS (CISA)	<ul style="list-style-type: none"> <li>Develop guidelines and best practices – NIST to focus on safety and trustworthiness of AI while CISA’s main role is to develop a comprehensive roadmap for AI initiatives as well as to focus on critical infrastructure.</li> <li>Coordinate interagency efforts led by CISA for national AI strategy which involves NIST drafts and guidance.</li> <li>Develop red teaming and exercises for AI safety and assessments.</li> </ul>
Executive Office of the President and DOD	<ul style="list-style-type: none"> <li>Coordinate federal AI usage by convening an interagency council to oversee AI development and implementation across government agencies.</li> <li>Develop guidelines for security reviews of federal data, balancing public access with the need to manage risks related to CBRN weapons and autonomous cyber capabilities.</li> </ul>
DOE and the OSTP (within the Executive Office)	<ul style="list-style-type: none"> <li>Coordinate to leverage AI for climate resilience, clean energy deployment, and grid reliability, while advancing partnerships with industry and academia to develop AI tools that enhance environmental and social outcomes.</li> <li>Lead initiatives to harness the Department of Energy’s AI capabilities for science, energy, and national security applications.</li> </ul>

Notably, the EO emphasized interagency coordination and consultation rather than establishing a new AI-specific agency. This approach fosters synergies among federal agencies and departments, allowing them to leverage their unique expertise in addressing AI challenges. However, despite this collaborative “whole of government” effort, the absence of a centralized AI agency means that agencies and organizations may still address AI risks independently, potentially leading to varied approaches.

## Impacts & Implications

Following the signing of EO 14110, the Biden administration was widely praised for placing values, ethics, and democratic principles at the center of its governance approach to AI.<sup>156</sup> The impacts of the AI EO can be summarized into four key observations<sup>157</sup>:

- **Developing a US vision of AI governance:** The Biden administration has begun to develop a US vision of AI governance grounded in democratic principles and laid out a rough template that Congress can adapt. This approach also signals to US companies and the international community that the US governance approach will focus on the relationship between AI and democratic health.
- **Focus on mitigating AI harms:** The administration’s focus squarely on harms is an acknowledgment that developments in AI have the potential to fundamentally reshape societies and economies.
- **Incorporating a hybrid approach:** The EO builds on foundations laid by the Blueprint for an AI Bill of Rights and NIST’s AI RMF which adopt rights- and risk-based governance models, respectively. This approach demonstrates to Congress that the goal is to embrace both the human-centric, rights-based approach as well as a product safety approach.
- **Underscoring public participation in AI governance:** The AI EO integrates civil society feedback from the Blueprint for an AI Bill of Rights (detailed below), which focuses on protecting vulnerable groups from AI threats. While the EO includes elements from company commitments made at the White House, it emphasizes meaningful public engagement by soliciting comments on the OMB’s draft guidance. This approach ensures that public perspectives influence the EO’s implementation.

As the 2024 US presidential election is around the corner, there are questions concerning the potential impact to the AI EO if Donald Trump were to win the election. Given the nature of presidential election campaigns and Donald Trump’s character, it is not entirely clear what Trump’s stance on AI regulation is nor what he plans to put into legislation if he were to win office. However, previously, Trump has described Biden’s AI EO as “dangerous,” claiming that Republicans support AI development rooted in free speech and human flourishing.<sup>158</sup> Additionally, in a 2023 campaign rally, Trump has pledged to repeal Biden’s AI

EO and “ban the use of AI to censor the speech of American citizens on day one.”<sup>159</sup> There is no clear indication that restrictive measures like mandatory testing will be completely abandoned, but Trump’s push for freedom of speech may influence his AI.

## Harris, Trump on key tech issues

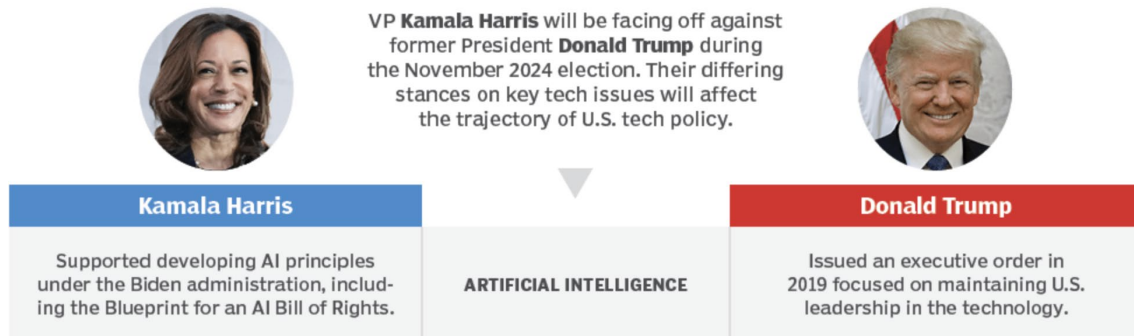


Figure 9: Comparison of 2024 Presidential Candidates’ Views on AI<sup>160</sup>

However, Biden and Trump have expressed similar views on AI regulation as well. Both support reducing regulatory barriers to foster innovation and growth and mitigating China’s dominance in the AI industry. Furthermore, Trump’s strategy towards AI while he was in office from 2016 to 2020 was similar to Biden’s approach<sup>161</sup>:

- 2019: Trump supported OECD’s 2019 AI principles.<sup>162</sup>
- February 2019: Trump signed an executive order to sustain American leadership in AI, launching the American AI Initiative.<sup>163</sup>
- 2020: Trump’s administration co-founded the Global Partnership on AI.<sup>164</sup>
- February 2020:<sup>165</sup> Trump committed to doubling nondefense AI research and development funding over two years.<sup>166</sup>
- December 2020: Trump signed an executive order to promote the use of AI within the federal government.<sup>167</sup>

Trump’s verbal support<sup>168</sup> for less AI restrictions may be favorable to GenAI developers, and Trump’s general leniency on climate policies—in relation to his decision on the Paris climate Accord—is favorable to GenAI developers as well.<sup>169</sup> Currently, Trump’s allies are drafting an AI EO to launch “Manhattan Projects” for military technology and swift review of regulations, signaling a pro-Silicon Valley approach in a potential second Trump administration.<sup>170</sup>

## II. The Office of Science and Technology Policy (OSTP) Blueprint for AI Bill of Rights (AIBoR)

### Overview

The White House Office of Science and Technology Policy (OSTP) announced the Blueprint for an AI Bill of Rights (AIBoR) in October 2022 to outline principles for protecting the American public and guiding the ethical use of technology.<sup>171</sup> The principles are non-regulatory and non-binding: a "Blueprint," as advertised, and not yet an enforceable "Bill of Rights" with legislative protections.<sup>172</sup> The AIBoR includes many examples of AI use cases that the White House OSTP considers problematic. Importantly, the document clarifies that the Blueprint should only apply to automated systems that have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services, generally excluding many industrial and/or operational applications of AI.<sup>173</sup> The AIBoR expands on examples of the use of AI in Lending, Human Resources, surveillance, and other areas (which are also covered by the "high-risk" use case framework of the EU AI Act).<sup>174</sup> The purpose of the AIBoR is to "help guide the design, use, and deployment of automated systems to protect the American Public."<sup>175</sup> In doing so, the blueprint identifies five principles including<sup>176</sup>:

- **Safe and Effective Systems:** Ensure systems are safe and effective and mitigate risks.
- **Algorithmic Discrimination Protections:** Prevent unjust treatment based on protected characteristics.
- **Data Privacy:** Protect privacy through design choices and user control over data.
- **Notice and Explanation:** Clearly explain automated system outcomes and usage.
- **Human Alternatives, Consideration, & Fallback:** Allow opt-out options and accessible recourse to human oversight.

The AIBoR defines an "automated system" as "any system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities."<sup>177</sup> Examples of such automated systems include real-time facial recognition systems, social media monitoring, systems that use or collect health-related data, ad-targeting systems, admissions algorithms, hiring or termination algorithms, and loan allocation algorithms. The framework outlines protections for all automated systems that could affect:

- **Civil Rights, civil liberties, and privacy,** including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both public and private sector contexts.
- **Equal opportunities,** including equitable access to education, housing, credit, employment, and other programs.
- **Access to critical resources or services,** including healthcare, financial services, safety,

social services, non-deceptive information about goods and services, and government benefits.

## Role of Agencies and Departments

As the AIBoR was announced, the Biden administration announced various actions across the federal government that have sought to advance the Blueprint, as described below.

Table 15: Overview of Actions by Agencies in Advancing the Blueprint for an AIBoR<sup>178</sup>

Agency / Initiative	Action
Department of Labor	Released “What the Blueprint for an AI Bill of Rights Means for Workers” and increased enforcement of surveillance reporting to protect workers. <sup>179</sup>
DOJ and the Equal Employment Opportunity Commission (EEOC)	Released antidiscrimination technical assistance and guidance on the Americans with Disabilities Act (ADA) and employment algorithms <sup>180</sup> ; launched a multi-year effort to improve hiring and recruitment practices using automated systems. Released antidiscrimination guidance on employment algorithms with the EEOC. <sup>181</sup>
EEOC and the Department of Labor	Launched a multi-year effort to reimagine hiring and recruitment practices, including automated systems. <sup>182</sup>
Consumer Financial Protection Bureau (CFPB)	CFPB affirmed that federal anti-discrimination laws mandate creditors to give clear and precise reasons when denying credit applications or taking adverse actions, regardless of the use of complex, black-box credit models. <sup>183</sup> The CFPB is intensifying efforts against algorithmic discrimination and expanding its team with technologists to enhance oversight.
Department of Education	Released AI guidelines for teaching and learning, focusing on safety, fairness, efficacy, and privacy. <sup>184</sup>
Department of HHS	Proposed rule to prohibit discrimination by algorithms in clinical decision-making, will examine on health care algorithms and disparities. <sup>185</sup> Requested information on mitigating bias in Medicare policy and algorithms.
Department of Veterans Affairs (VA)	Instituted a principle-based ethics framework for access to and use of veteran data and launched AI@VA to manage AI risks in healthcare.
Department of Housing and Urban Development (HUD)	Released guidance on tenant screening algorithms and their compliance with the Fair Housing Act. <sup>186</sup>
United States Agency for International Development (USAID)	Launched an AI Action Plan to embed risk mitigation in AI and support responsible technology worldwide. <sup>187</sup>
OMB, OSTP, Federal Chief Information Officers Council	Coordinated across the government to publish inventories of non-classified government AI use cases to adhere with civil rights and privacy laws. <sup>188</sup>
DOE	Released Principles and Guidelines for Responsible and Trustworthy AI and an AI Risk Management Playbook. <sup>189</sup>
DOD	Operates under AI Ethical Principles and a Responsible AI Strategy & Implementation Pathway. <sup>190</sup>
Intelligence Community (IC)	Operates under Principles of AI Ethics and an AI Ethics Framework. <sup>191</sup>
NSF	Invests over \$700 million annually in AI research, focusing on fairness, security, safety, and trustworthiness.

## Impacts & Implications

The five principles and associated practices of the AIBoR create a comprehensive framework to safeguard against potential AI threats. While experts assess that the Blueprint reflects a significant step towards addressing the challenges posed by AI, its impact and implications reveal both progress and shortcomings in US AI regulation:

### Progress<sup>192</sup>

- **Sector-specific guidance:** the AIBoR details actions that address most high-priority algorithmic harms across healthcare, financial services, education, and housing.
- **Foundation to build capacity:** the AIBoR covers a wide range of issues and federal actions that can be leveraged to expand capacity in the future and provides a foundation for future AI regulation in the US and internationally.

### Shortcomings<sup>193</sup>

- **Uneven Progress on Algorithmic Protections:** the Blueprint focuses primarily on sectors like financial services and healthcare. This targeted approach has led to demonstrable progress in these areas, such as the Federal Trade Commission (FTC)'s proposed rules on commercial surveillance and the CFPB's requirements for explanations in credit denials. However, sectors like education, workplace surveillance, and law enforcement have insufficient regulatory attention.
- **Implementation Challenges:** while several federal agencies have begun addressing AI-related issues, some agencies have not responded adequately to AI governance challenges. For instance, the Department of Labor limited its focus to surveillance related to labor organizing, neglecting broader employee surveillance concerns.
- **Lack of Binding Guidance:** the AIBoR provides nonbinding principles, which limits its immediate impact. The effectiveness of these principles largely depends on the actions of federal agencies rather than having a direct regulatory force.
- **Missed Coordination Opportunities:** The White House has been criticized for not effectively coordinating and facilitating AI regulation across agencies to address common challenges and barriers in AI governance. There is no clear public evidence of coordination between AIBoR, AI EO, and other governance approaches which could cause inefficiency and confusion.

Though the AIBoR promoted initial discussions and actions on AI regulation, experts suggest its uneven implementation and the limited scope of federal agency responses highlight the need for ongoing attention and refinement in AI governance.

### III. National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF): Generative AI Profiles

#### Overview

NIST released the AI RMF 1.0 in January 2023 and highlighted voluntary adoption. The RMF aimed to integrate considerations of trustworthiness into the design, development, use, and evaluation of AI products, services, and systems. The framework was developed through a collaborative and transparent process which included a Request for Information, multiple drafts open for public comment, various workshops, and other input opportunities. Interview findings note that NIST recognized similarities in previous approaches to cybersecurity and risk management in creating the AI RMF.<sup>xxvii</sup>

Under the AI EO, NIST was tasked with a leading role in implementing many of the directives. In April 2024, NIST released a companion document—the NIST AI 600-1 AI RMF Generative AI Profile<sup>194</sup>—which serves as a use-case and cross-sectoral profile of the AI RMF 1.0.<sup>195</sup> Use-case profiles apply the AI RMF functions, categories, and subcategories to specific settings or applications such as GenAI. These profiles are tailored to the requirements, risk tolerance, and resources of the Framework user. Like other AI RMF Profiles, this profile provides guidance on managing risks throughout different stages of the AI lifecycle. Similarly, cross-sectoral profiles are designed to govern, map, measure, and manage risks associated with activities or business processes that are common across different sectors, such as the use of LLMs, cloud-based services, and acquisitions.

The draft profile outlines risks unique to or intensified by generative AI, offering key actions for governance, mapping, measurement, and management. Key risks identified include<sup>196</sup>:

- CBRN Weapons Information: Risks associated with chemical, biological, radiological, or nuclear weapons data.
- Confabulation: Issues like “hallucinations” or “fabrications” in GenAI outputs.
- Dangerous Recommendations: Potential for GenAI to produce harmful or violent suggestions.
- Data Privacy: Concerns regarding sensitive data such as biometrics, health, location, and personally identifiable information.
- Environmental Impact: Resource use in training GenAI models.
- Human-AI Interaction: Risks from the interaction between humans and AI, such as “algorithmic aversion,” automation bias, or misaligned goals.
- Information Integrity: Ensure accuracy and reliability of GenAI created information.
- Information Security: Protection of data and information security.

---

<sup>xxvii</sup> Data Consultant 1 interview.



- Intellectual Property: Risks related to intellectual property management.
- Obscene Content: Issues with obscene, degrading, or abusive content.
- Toxicity and Bias: Risks of toxicity, bias, and homogenization in GenAI outputs.
- Value Chain Integration: Challenges with non-transparent or untraceable integration of third-party components, including data acquisition and supplier vetting.

The AI RMF also aligns to a certain degree with the AIBoR as the AIBoR was a foundation to build capacity and covered a wide range of federal actions. Four of AIBoR's principles overlap with AI RMF's key issues:

- AIBoR's principle for safe and effective system is parallel to AI RMF's risks of confabulation and dangerous recommendations.
- AIBoR's principle for algorithmic discrimination protections is parallel to AI RMF's risk of toxicity and bias.
- AI AIBoR's principle for data privacy aligns with AI RMF's risk of data privacy.
- AI BoR's principle for notice and explanation is parallel to AI RMF's risk of value chain integration.

In addition to the AI RMF document, US NIST also released several "Crosswalk Documents," mapping concepts and terms between the AI RMF 1.0 and various guidelines, frameworks, standards, and regulatory documents. One crosswalk compares the NIST AI RMF 1.0 to the Japan AI Guidelines for Business (AI GfB) and notes several similarities and differences in terminology.<sup>197</sup> The comparison is outlined in Appendix C.

## Impacts & Implications

The NIST AI RMF has received both positive and negative reactions from experts. On the positive side, experts note the following impacts of the framework<sup>198</sup>:

- **Guidance in the absence of federal legislation:** the NIST AI RMF plays an important role in guiding AI development and governance in the US, emphasizing the protection of individual rights and privacy.
- **Timely recognition of risk management practices:** the NIST AI RMF was ahead of its time by identifying that managing AI risks closely aligns with established practices for other applications, rather than creating new approaches.
- **Flexibility and applicability:** the framework is robust due to its flexibility and relevance across diverse use cases, acknowledging that AI risk management is highly context dependent.
- **Comprehensive governance:** the framework addresses both the development and usage of AI systems, providing more comprehensive guidance and a multi-stakeholder approach compared to some other proposed frameworks.

- **Integration with existing knowledge:** the NIST AI RMF builds on existing privacy and security knowledge from established frameworks and global standards, particularly around cybersecurity, avoiding the need to “reinvent the wheel” for AI.
- **International relevance:** The NIST AI RMF resonates with global efforts, aligning with frameworks from the EU, Singapore, and the OECD, enhancing its global relevance.
- **Support for existing initiatives:** the framework complements the Blueprint for an AI Bill of Rights and federal guidance on algorithmic discrimination, providing a practical approach for implementing these guidelines across various sectors.

However, other commentary points to the negative impacts and implications of the framework<sup>199</sup>:

- **Lack of binding authority:** the voluntary nature of the NIST AI RMF may limit its enforcement and impact.
- **Technical complexity:** the AI RMF’s technical nature may challenge those not deeply familiar with AI risk management potentially limiting its accessibility to policymakers.
- **Adoption challenges:** despite its flexibility, the AI RMF’s non-binding nature does not guarantee widespread adoption and could face resistance from organizations preferring more concrete regulations.
- **Broad scope:** the framework’s broad and adaptable approach might lack the precision needed for potential detailed regulation, potentially leading to challenges in addressing specific AI applications or risks effectively.

#### IV. Other NIST Drafts

##### Overview

In addition to the AI RMF 1.0, NIST has released several other AI-related drafts in response to the AI EO, including:

- NIST Secure Software Development Framework (SSDF) for Generative AI and Dual-Use Foundation Models: SP800-218A (SSDF profile) released in July 2024.
- NIST Plan for Global Engagement on AI Standards (NIST AI 100-5) released in July 2024.
- NIST AI100-4 Reducing Risks Posed by Synthetic Content released in April 2024.

NIST Secure Software Development Framework (SSDF) for GenAI: SP800-218A (SSDF profile)  
 President Biden’s AI EO tasked NIST with “developing a companion resource to the Secure Software Development Framework (SSDF) to incorporate secure development practices for generative AI and dual-use foundation models.”<sup>200</sup> SSDF is based on secure software

development practices from organizations like the Software Alliance<sup>xxviii</sup> (known as BSA), the Open Worldwide Application Security Project<sup>xxix</sup> (OWASP), and SAFECode.<sup>xxx</sup> It aims to enhance software security within the software development life cycle (SDLC) by integrating with existing SDLC models.<sup>201</sup>

The objectives of SSDF include reducing vulnerabilities in released software, mitigating the impact of exploited vulnerabilities, and addressing root causes to prevent future vulnerabilities.<sup>202</sup> The SSDF focuses on the various phases of AI model development, spanning from data sourcing and training to software integration. However, it does not address the deployment or operation of AI systems, nor does it cover broader data governance aspects outside of cybersecurity practices for training data. The SSDF was recently released for public comment until June 2024. The SSDF outlines several key practices to enhance secure software development, organized into four groups<sup>203</sup>:

- Preparing organization (PO) to ensure readiness for secure software development
- Protecting software (PS) by safeguarding all components from unauthorized access
- Producing well-secured software (PW) to minimize security vulnerabilities in releases
- Responding to vulnerabilities (RV) by identifying and addressing any residual vulnerabilities effectively

In two of the groups above, the SSDF for GenAI specifies recommendations for training data. PO recommendations include having artifacts that include the attestations of training data integrity and provenance as well as continuously monitoring AI-related resources which include training data. PW requires the confirmation of training, testing, and fine-tuning data before model usage and recommends including AI model-specific threat types in risk monitoring, including poisoning training data.

Furthermore, the SSDF is intended to align secure software development with business and mission requirements, ensuring that it meets organizational goals, risk tolerance, and available resources. It provides actionable insights and prioritization by comparing current outcomes with SSDF practices. This comparison helps identify gaps and guide the development of prioritized action plans based on the organization's mission and risk management strategies.

---

<sup>xxviii</sup> The Software Alliance (BSA) is a trade group of business software companies established in 1998. Its principal activity is trying to stop copyright infringement of software produced by its members.

<sup>xxix</sup> The OWASP is an online community that produces freely available articles, methodologies, documentation, tools, and technologies in the fields of IoT, system software and web application security.

<sup>xxx</sup> SAFECode is a global nonprofit organization that brings business leaders and technical experts together to exchange insights on creating, improving and promoting scalable and effective software security programs.

### NIST Plan for Global Engagement on AI Standards (NIST AI 100-5)

The AI EO directs the DOC to devise a plan for global engagement—Sections 11(b), 11(b)(i), 11(b)(ii), 11(b)(iii)—which includes developing tools for implementing standards and promoting cross-sectoral standards. In response, NIST’s Plan for Global Engagement on AI Standards outlines a collaborative effort with international allies, partners, and standards organizations to create and implement consensus standards for AI.<sup>204</sup> It seeks to engage a range of global experts from various disciplines, ensuring alignment with US standards and interests. The Plan’s goals include:

- Creating standards that are accessible and easy to adopt
- Reflecting diverse global stakeholder needs
- Ensuring an open, transparent, and consensus-driven development process
- Strengthening international partnerships

NIST identifies three priority categories for standardization: 1) Urgently needed and ready for standardization; 2) Needed but requiring further scientific research; 3) Needed but requiring significant foundational work.

### NIST AI 100-4 Reducing Risks Posed by Synthetic Content

NIST AI 100-4 addresses the potential harms and risks associated with AI-created or altered content. It offers guidance on detecting, authenticating, and labeling synthetic content.<sup>205</sup> The document delves into methods such as digital watermarking<sup>xxxii</sup>, metadata recording<sup>xxxiii</sup>, and strategies for identifying AI-generated images, videos, text, and audio. The focus includes essential aspects like authenticating and tracking content provenance, labeling synthetic content, detecting harmful materials—including child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII)—and ensuring transparency through testing and maintenance of synthetic content.

Drawing from public feedback and consultations, the report underscores the importance of digital content transparency in building trustworthiness. Additionally, the report brings attention to ongoing global efforts to develop scientifically backed standards for these tools, emphasizing the role of digital literacy in fostering public trust.

---

<sup>xxxii</sup> Digital watermarking refers to the method of embedding unique identifiers into digital content, such as images, videos, or documents, to protect intellectual property and verify authenticity.

<sup>xxxiii</sup> Metadata can provide information about a set of data, its origin, and its content and contribute to digital content transparency. Metadata can be generated whenever digital content is created, uploaded, downloaded, or modified.

## Impacts & Implications

NIST's AI-related drafts emphasize trustworthiness, transparency, and international standardization to protect training data and AI models. While the drafts remain voluntary guidelines, each is notable in driving secure and responsible AI development, including:

- **Enhanced Security in the AI Model Lifecycle:** the SSDF integrates secure development practices—including protecting training data against risks—into the AI model lifecycle, aiming to reduce vulnerabilities and mitigate the impact of potential security breaches. Although the SSDF focuses on pre-deployment phases, its practices could influence broader software security.
- **Strengthened Global Standards:** NIST's Plan for Global Engagement on AI Standards aims to create and adopt standards that align with US interests, positioning the US as a global leader in AI safety. This framework could enhance global standardization and interoperability of AI systems, setting a precedent for international regulations.
- **Improved Detection and Authentication of Synthetic Content:** NIST's publication on synthetic content outlines clear methods for identifying and managing AI-generated content which are crucial for building public trust and mitigating the increasing risks associated with synthetic media. As the risks of synthetic media continue to threaten election campaigns, NIST's work provides a proactive approach to election security.
- **Trust and Transparency:** The publications outline initiatives to enhance the credibility of AI-generated content and support informed consumer interactions, broadly contributing to improving AI digital literacy and transparency.

## V. Information technology - AI – Management system (ISO/IEC 42001:2023)

### Overview

ISO/IEC 42001:2023 introduces a new framework in AI Management Systems (AIMS).<sup>206</sup> It outlines the requirements for establishing and maintaining an AI management system within organizations, focusing on issues such as opaque decision-making and the adaptive learning capabilities of AI systems. Organizations are encouraged to incorporate AI management into their existing frameworks, considering factors like organizational goals, stakeholder expectations, and customized risk management approaches for specific AI applications. Through a “Plan-Do-Check-Act” methodology, the standard stresses the importance of embedding AI-specific concerns into organizational processes, such as risk, lifecycle, and supplier management, to ensure responsible and accountable use of AI technologies.<sup>207</sup> The standard specifies requirements for AIMS that include<sup>208</sup>:

- Policies and objectives for responsible AI development and use.
- Processes to achieve these objectives.

- Ethical considerations, transparency, and accountability in AI systems.
- A “Plan-Do-Check-Act” methodology for managing AI-related risks and opportunities.
- Performance measurement, including both quantitative and qualitative outcomes.
- Conformity to requirements and systematic audits to assess AI systems.

ISO/IEC 42001:2023 outlines key processes for effective AI governance, including:

- Risk Management: Identifying and addressing AI-specific risks.
- Data Quality and Governance: Ensuring high standards for data management.
- Policy and Accountability: Establishing clear policies and accountability structures.
- Continuous Improvement: Regularly evaluating and enhancing AI system performance.
- Documentation & Rationalization: Maintaining records of AI controls and decisions.

The standard also offers detailed guidance on AI system development. Annex A provides a comprehensive list of controls while Annex B focuses on data management processes. Additionally, the standard helps organizations identify AI-related objectives and risks and offers insights into specific domain and sector standards.

### Impacts & Implications

The ISO/IEC 42001:2023 standard significantly influences organizations' AI practices across various aspects, including:

- **Structured Management of AI Systems:** provides a framework for ethical AI development and ongoing improvement, helping organizations stay current and compliant with best practices.
- **Ethical and Transparent AI Development:** enhances alignment with ethical standards and human rights, boosting public trust and reducing the risk of ethical breaches.
- **Risk and Impact Management:** encourages systematic risk assessment and impact evaluation, mitigating potential harms and legal issues.
- **Data Quality and Governance:** ensures data quality and quality of AI to increase the security, safety, and fairness of systems and information while adhering to data protection laws.
- **Organizational Credibility and Reputation:** offers independent validation of AI practices, enhancing credibility and reputation, and potentially providing a competitive edge.

## VI. Information technology - AI – Guidance on risk management (ISO/IEC 23894:2023)

### Overview

ISO/IEC 23894:2023 aligns with ISO 31000:2018<sup>xxxiii</sup>, extending risk management guidance with specific AI considerations where applicable. This standard is divided into three main sections:<sup>209</sup>

- **Clause 4 Principles:** This section outlines foundational principles of risk management adapted to address AI-specific considerations. It provides a comprehensive overview of the core principles that should guide the management of AI-related risks.
- **Clause 5 Framework:** The purpose of the risk management framework is to assist organizations in integrating risk management into key activities associated with AI development, provisioning, and use. It offers practical guidance on embedding risk management practices into significant AI-related processes.
- **Clause 6 Processes:** This section details systematic risk management procedures tailored specifically for AI applications. It outlines structured approaches for identifying, assessing, and managing risks throughout the AI lifecycle.

The standard also includes several annexes:

- **Annexes A and B:** These annexes identify common objectives and risk sources related to AI. They offer insight into typical risks associated with AI systems, such as algorithmic biases and data privacy, and guide on addressing these risks effectively.
- **Annex C:** This annex illustrates how risk management processes align with the AI system lifecycle, highlighting how risk management practices can be integrated throughout the various stages of AI development and deployment.

ISO/IEC 23894:2023 aims to provide a structured approach to managing AI-related risks, ensuring that organizations can apply comprehensive and specialized risk management strategies in their AI operations.

### Impacts & Implications

ISO/IEC 23894:2023 has been praised for its focus on addressing algorithmic biases, providing comprehensive guidelines to tackle these issues such as advising organizations to evaluate training data for historical biases, using diverse datasets and regularly testing models for fairness across different demographic groups.<sup>210</sup>

---

<sup>xxxiii</sup> ISO 31000:2018 is an international standard that provides principles and guidelines for risk management. It outlines a comprehensive approach to identifying, analyzing, evaluating, treating, monitoring and communicating risks across an organization.

Furthermore, the standard also seeks to address the “black box” problem, whereby AI models are so complex that even their developers struggle to explain their decisions. The standard emphasizes the importance of explainable AI (XAI), recommending the use of interpretable models and techniques like LIME<sup>xxxiv</sup> (Local Interpretable Model-agnostic Explanations) or SHAP<sup>xxxv</sup> (SHapley Additive exPlanations).<sup>211</sup> This focus on transparency could prevent issues like the healthcare algorithm's underestimation of Black patients' needs, promoting more accurate and equitable AI applications.<sup>212</sup>

ISO/IEC 23894 addresses data privacy risks by promoting techniques such as differential privacy<sup>xxxvi</sup> and federated learning.<sup>xxxvii</sup> This guidance is crucial in preventing privacy breaches like the Cambridge Analytica scandal, ensuring that organizations only collect and retain necessary data while protecting user privacy.<sup>213</sup>

Importantly, the standard’s guidelines for autonomous systems, including formal verification and scenario-based testing, aim to enhance safety in high-risk applications like autonomous vehicles.<sup>214</sup> It also highlights the importance of establishing clear liability frameworks to determine responsibility in case of an accident. The NIST AI RMF along with NIST SSDF and ISO/IEC 23894:2023 serve complementary yet distinct roles in AI risk management. While the NIST AI RMF emphasizes adaptability and community engagement for practical risk management throughout the AI lifecycle and the NIST SSDF focuses on the risks seen in the development lifecycle, the ISO/IEC 23894:2023 promotes global consistency in AI risk management, focusing on assessment, treatment, and transparency.

## VII. State Regulations

### Overview

AI regulations at the US state level are currently fragmented and evolving, with various states introducing and advancing regulatory frameworks tailored to their specific concerns. In the 2024 legislative session, a significant number of jurisdictions, including at least 40 states, Puerto Rico, the Virgin Islands, and Washington, D.C., introduced AI-related bills.<sup>215</sup> Of these, seven states and Puerto Rico enacted resolutions or legislation addressing AI.

---

<sup>xxxiv</sup> LIME is a technique that approximates any black box machine learning model with a local, interpretable model to explain each individual prediction.

<sup>xxxv</sup> Shapley Additive Explanations (SHAP) is a game theory-based method for explaining the output of machine learning models. SHAP values are used to show how much a feature or input contributes to a model's prediction, and how each feature affects the final prediction.

<sup>xxxvi</sup> Differential privacy is a mathematical definition ensuring that the output of an algorithm analyzing a dataset does not reveal whether any individual's data was included, by maintaining almost identical behavior whether or not a single individual's data is present in the dataset.

<sup>xxxvii</sup> Federated learning (also known as collaborative learning) is a sub-field of machine learning focusing on settings in which multiple entities (often referred to as clients) collaboratively train a model while ensuring that their data remains decentralized.



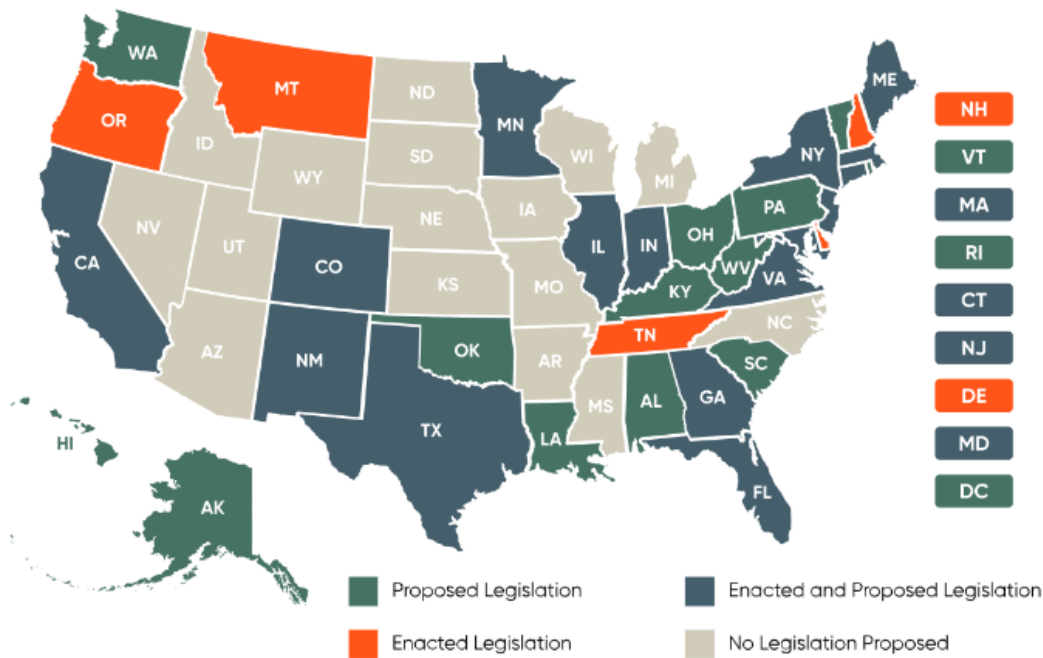


Figure 10: Overview of State-by-State AI legislation<sup>216</sup>

Recent examples of state-level AI regulations include<sup>217</sup>:

- **Colorado:** Enacted measures requiring developers and deployers of high-risk AI systems to exercise reasonable care to prevent algorithmic discrimination. Additionally, the state mandated disclosures to consumers about AI systems.
- **Florida:** Allocated grants to school districts for implementing AI technologies to support students and teachers, highlighting a focus on educational applications of AI.
- **Indiana:** Established a task force to guide AI initiatives and developments in Indiana.
- **Maryland:** Adopted procedures governing the development, procurement, deployment, use, and assessment of AI systems by state government units.
- **South Dakota:** Revised its laws to clarify that possessing child pornography includes visual depictions or simulations involving minors, including computer-generated content. Violations of this updated law are classified as a Class 4 felony.
- **Tennessee:** Mandated that governing boards of public institutions of higher education create rules regarding AI usage. Local education boards and public charter schools are also required to adopt policies for AI use by students, teachers, faculty, and staff for instructional purposes.
- **Utah:** Created the AI Policy Act to address some aspects of AI regulation within Utah.
- **West Virginia:** Formed a select committee on AI to focus on AI-related issues and advise on state-level policies and regulations.

Additionally, California is a leading state in developing and passing data, privacy, security, and technology related bills. Often, national legislation follows those of state regulations and

especially that of California. California's state legislature proposed stringent legislation that included compliance restrictions for large and powerful AI models, civil and criminal liability for developers, 30 new AI-related measures, and more. At the end of September 2024, California Governor Newsom vetoed the bill, perpetuating the debate regarding the appropriate balance between AI safety and innovation. <sup>218</sup> Appendix D provides further information on selected states' AI legislation and the categorization of their approach.

## Impacts & Implications

Overall, states are increasingly implementing targeted legislation to address specific concerns, as comprehensive federal regulation remains absent. Given the varying priorities and stringency of legislation across states, the impacts will differ regionally. However, the growing trend towards state-level regulation reflects several common themes:

- **Increased Accountability and Compliance Costs:** Many states are introducing detailed regulations that require businesses to adopt transparency, fairness, and accountability measures. This includes compliance requirements for high-risk AI applications, such as mandatory disclosures, bias audits, and human review processes. As a result, companies may face higher operational costs and legal risks.
- **Enhanced Consumer Protection:** Legislation often aims to shield consumers from potential AI-related harms (discrimination, privacy breaches, and misinformation).
- **Combatting Discriminatory Hiring Practices:** Several states are implementing regulations focused on AI in hiring, such as mandatory bias audits and transparency requirements for automated decision-making tools used in employment.
- **Promotion of Ethical AI Practices:** By setting standards for transparency and ethical use, state regulations encourage the development of responsible AI technologies. Rules on deepfakes, facial recognition, and political advertising emphasize the importance of ethical considerations in AI deployment and content creation and can emphasize legislation at the federal level.
- **Regulatory Burdens and Investment:** States with more stringent AI regulations, such as those mandating extensive bias audits, transparency measures, and "kill switch" requirements, may present higher compliance costs and operational hurdles for businesses. As a result, AI startups and established companies might be deterred from investing in or expanding within these states, leading to slower innovation and development in those regions.

## C. Summary

Collectively examining the US AI governance approaches, it is clear that the US aims to mitigate and safeguard against the five AI-enabled risks reported in the first report. At a high-level, preexisting cybersecurity and safety approaches can help mitigate against AI-enhanced

threats. The AI EO, the AIBoR, and the ISO/IEC 42001:2023 promote ethical use of AI, putting some responsibility on the technology users as well. The AI EO and the AIBoR also aim to increase workforce AI literacy and public awareness of AI risks which can help empower the people to safeguard against threats. Finally, all of the highlighted US governance approaches attempt to mitigate AI threats at varying levels of the AI development and deployment lifecycle, creating multiple layers of protection. Additional analysis on the strengths and weaknesses of EO 14110, NIST AI RMF, NIST SSDF for GenAI, AIBoR, and CA SB-1047 can be found in Appendix E.

Furthermore, when mapping the five categories of AI-enabled risks to the US AI governance approaches, the approaches lend themselves to cover all five categories to some degree:

Table 16: Coverage of AI-Enable Risks by US Governance Approaches

Threat / Governance Approach	EO 14110	AI BoR	NIST RMF	NIST SSDF for GenAI
AI-Enhanced Traditional Cyberattacks	Well covered—emphasizes cybersecurity, including measures for offensive cyber operations and guidelines for auditing AI capabilities to mitigate potential harm.	Adequately covered—aims for safe and secure systems which would mitigate against traditional cyberattacks.	Not covered.	Well covered—Focuses on protecting software from unauthorized access and producing well-secured software with minimized vulnerabilities.
AI-Enabled Disinformation & Misinformation	Well covered—focus on safe and secure systems as well as defending against related cyberattacks can defend against false information, requires DOC to work on authenticating GenAI content.	Adequately covered—suggests human alternatives and fallbacks which can filter false information.	Well covered—Confabulation, toxicity and bias, and obscene content risks.	Adequately covered—by ensuring a safe and secure AI development lifecycle, hallucinations can be mitigated which will lead to less misinformation.

Threat / Governance Approach	EO 14110	AI BoR	NIST RMF	NIST SSDF for GenAI
AI-Enabled Disruption or Maloperation of Systems	Well covered—requires agencies to coordinate to develop guidelines that ensure critical infrastructure systems’ resiliency suggests red teaming and other assessments to ensure quality AI systems, and more.	Adequately covered—algorithmic discrimination protection, human alternatives and fallbacks, and safe and effective systems can help mitigate system disruptions.	Adequately covered—identifies data integrity, information security, and dangerous recommendation risks.	Adequately covered—by ensuring a safe and secure AI development lifecycle, disruption and maloperation of systems can be mitigated.
AI-Enabled National Security Threats	Well covered—AI EO is the US’s foundational strategy for AI which is largely aimed at protecting the nation from CBRN risks to critical infrastructure risks to more.	Adequately covered—safe and secure systems can mitigate national security risks, and civil rights/liberties and privacy can help mitigate against espionage.	Adequately covered—identifies CBRN weapons risk.	Adequately covered—by ensuring a safe and secure AI development life cycle, eventually impacts national security.
Business Risks Due to Misuse of GenAI	Well covered—much of the DOC’s work is related to securing businesses from potential GenAI risks.	Well covered—allowing users access to equal opportunities and resources, protecting user data and privacy, mitigating against algorithmic discrimination, and requiring safe and secure systems can help reduce business risks.	Well covered—identifies data privacy, intellectual property, dangerous recommendations, value chain integration, and Human-AI interaction risks.	Well covered—refers to risks in the AI development lifecycle from preparing the organization to responding to vulnerabilities.

While it is evident that US AI governance approaches collectively address AI-enabled risks, the relative recency of these AI governance measures and the rapid speed at which AI technologies evolve can influence the exact mitigation impact of current AI governance measures. Furthermore, the lack of comprehensive AI legislation in the US as well as enforcing mechanisms on the aforementioned approaches could lead to ineffective safeguarding against AI risks and threats that increase confusion.

## 7. Criteria for Effective AI Governance and Framework Measures

### A. Background

Rapid AI advancements have prompted debates about the most effective governance measures. Emerging governance approaches reflect the ongoing challenges introduced by GenAI technologies' scale, power, and design. Figure 11 below highlights key debates in AI governance, focusing on whether to prioritize long-term existential risks or address immediate AI harms, and the trade-offs between open-source and closed-source AI. While some argue that focusing on existential AI risks could prevent future disasters, others caution that this approach may divert resources from addressing current issues. The open-source versus closed-source debate weighs the benefits of innovation and accessibility against the risks of security vulnerabilities and misuse. These debates underscore the complexity of developing AI governance frameworks which must balance innovation, safety, and ethical considerations.

TABLE 2 | Areas of debate in AI governance (non-exhaustive)

Debate and context	Sample position	Policy arguments for	Policy arguments against
<b>Policy focus on long-term existential risks<sup>31</sup> vs present AI harms.<sup>32</sup></b> AI poses present harms and a spectrum of potential near- to long-term risks. Diverse positions exist regarding how to identify and prioritize the harms and risks from AI as well as the timeframe over which risks should be considered.	Advanced autonomous AI systems pose an existential threat to humanity. <sup>33</sup>	<ul style="list-style-type: none"> <li>Without sufficient caution, humans could irreversibly lose control of autonomous AI systems.<sup>34</sup></li> <li>Starting with the biggest questions around existential risk supports the development of trustworthy AI and could prevent overregulation.<sup>35</sup></li> </ul>	<ul style="list-style-type: none"> <li>Existential risks are speculative and uncertain.<sup>36</sup></li> <li>Can redirect the flow of valuable resources from scientifically studied present harms.<sup>37</sup></li> <li>Misdirects regulatory attention.<sup>38</sup></li> </ul>
	Effective regulation of AI needs grounded science that investigates present harms. <sup>39</sup>	<ul style="list-style-type: none"> <li>In terms of urgency, there are immediate problems and emerging vulnerabilities with AI that disproportionately impact marginalized and vulnerable populations.</li> <li>Contending with known harms will address long-term hypothetical risks.<sup>40</sup></li> </ul>	<ul style="list-style-type: none"> <li>Focus on known harms may lead to neglecting long-term risks not well considered by traditional policy goals.</li> </ul>
<b>Policy treatment of open-source vs closed-source AI.<sup>41</sup></b> Governance consideration is being given regarding where an AI technology may sit on a spectrum of open-to-closed access. <sup>42</sup>	Open-source AI is critical to AI adoption and mitigating current and future harms from AI systems. <sup>43</sup>	<ul style="list-style-type: none"> <li>Increased access to AI and democratization of its capabilities.</li> <li>Spurs innovation and stimulates competition.</li> <li>Enables study of risks that can reduce bias and disparate performance for marginalized populations.</li> </ul>	<ul style="list-style-type: none"> <li>Increased access exposes AI models to greater malicious use and unintentional misuse.</li> <li>Difficulties in patching vulnerabilities can leave the AI system unsecured.<sup>44</sup></li> </ul>
	Closed-source AI is necessary to protect against misuse of powerful AI technology. <sup>45</sup>	<ul style="list-style-type: none"> <li>Protects commercial intellectual property.</li> <li>Safeguards against potentially harmful future capabilities.</li> <li>Identified vulnerabilities can be fixed and safety features can be implemented.<sup>46</sup></li> </ul>	<ul style="list-style-type: none"> <li>Concentration of power and knowledge within high-resource organizations.<sup>47</sup></li> <li>Increased dependency on a few foundation model providers with the risk of monopoly-related consequences.</li> </ul>

Figure 11: Areas of debate in AI governance<sup>219</sup>

Other emerging debates involve the impact of GenAI on employment, its intersection with copyright laws, requirements for data transparency, and the distribution of responsibility among various stakeholders in the generative AI lifecycle.<sup>220</sup> Furthermore, the potential for

GenAI to amplify misinformation and disinformation presents serious challenges. Many of these issues stem from data governance concerns, such as privacy, data protection, embedded biases, and identity and security risks associated with both the data used to train generative AI systems and the data generated by these systems.<sup>221</sup>

This section examines the latest research on approaches to US AI governance, including insights from expert interviews on various governance strategies and methods for evaluating these measures. It explores how emerging debates influence the creation of a governance model and criteria to develop AI governance frameworks. The section concludes by analyzing how the derived criteria align with existing US governance frameworks.

## B. Findings from Research on US AI Governance Measures

Before the EO 14110, the US had a “laissez-faire approach to the governance of AI”, without a centralized federal regulatory framework dedicated to general-purpose AI.<sup>222</sup> Instead, AI regulation was fragmented, with various federal agencies independently developing and implementing new policies on AI, tailored to specific needs and contexts, but lacking a unified national strategy. This section explores existing research and debates on US AI governance measures, highlighting several key themes from discussions around industry self-governance, approaches to AI stacks (Section 7.B.II) and general-purpose AI, concerns surrounding intellectual property rights, and the call for establishing a separate AI agency.

### I. Industry Self-Governance

Despite the growing recognition of AI’s potential risks, comprehensive government regulation remains largely absent in the US, necessitating a closer look at industry self-governance as a viable alternative. Existing research on industry self-governance has identified the following issues and approaches<sup>223</sup>:

- **Industry vs. Organizational Self-Governance and Ethical AI:** Industry self-governance involves voluntary, collective actions by industry members to address societal concerns. This contrasts with organizational self-governance, where individual organizations establish their own policies and governance processes. Despite public declarations by many organizations of their adoption of trust-enhancing practices, there is significant divergence on what constitutes "ethical AI."
- **Evidence-based AI Risk Mitigation:** AI developers and implementers should more widely adopt evidence-based practices to mitigate risks. However, government regulation to enforce evidence-based mitigation practices is largely lacking.
- **Uncertainty in Governance Responsibility:** There is ongoing debate over whether the government or the private sector is best suited to manage AI risks and maintain public

trust. Industry self-governance may be necessary when government actions do not sufficiently address public concerns.

- **Multistakeholder Participation:** Effective self-governance efforts must involve a broad set of stakeholders, including consumers, AI developers, and government agencies, to ensure diverse perspectives are considered.
- **Operationalize Program Design:** Accreditation and certification programs need to be carefully designed. Accreditation might cover adherence to a comprehensive set of standards, while certification could be more targeted. The creation of market demand and the evaluation of these programs' effectiveness are key to ensuring that they contribute to responsible AI development.

When the industry engages in AI self-governance, it must include self-governance measures in all phases of the AI implementation cycle. Figure 12 outlines and summarizes key elements or standards for AI risk mitigation practices across different stages of AI implementation.

NAM Life cycle	Risks	Evidence-based practices
Phase 1: Needs Assessment	<ul style="list-style-type: none"> <li>• Lack of integration of stakeholder perspectives &amp; considerations<sup>16-22</sup></li> <li>• Lack of clearly defined organizational values &amp; ethics<sup>23,24</sup></li> </ul>	<ul style="list-style-type: none"> <li>• User-centered design<sup>25,26</sup></li> <li>• Organizational readiness assessment<sup>27-29</sup></li> <li>• Organizational prioritization process<sup>1</sup></li> <li>• User-centered workflow/change management process<sup>5,30-32</sup></li> </ul>
PHASE 2: Development	<ul style="list-style-type: none"> <li>• Data bias<sup>33-38</sup></li> <li>• Lack of representative &amp; equitable population<sup>33,39</sup></li> <li>• Lack of data management<sup>37,40</sup></li> <li>• No accounting for causal pathways<sup>41</sup></li> </ul>	<ul style="list-style-type: none"> <li>• Data transparency &amp; reporting<sup>32,37,40,42-46</sup></li> <li>• Model provenance records<sup>40</sup></li> <li>• Promoting trust &amp; explainability<sup>32,47-51</sup></li> <li>• Distributed model development<sup>52</sup></li> </ul>
PHASE 3: Implementation	<ul style="list-style-type: none"> <li>• Lack of data encryption &amp; privacy protections<sup>53,54</sup></li> <li>• Lack of secure hardware</li> <li>• Lack of oversight for responsible AI adoption<sup>39,55</sup></li> </ul>	<ul style="list-style-type: none"> <li>• Equitable/diverse workforce</li> <li>• Organizational implementation<sup>38,46,47,56,57</sup></li> <li>• Organizational governance<sup>13,47,58,59</sup></li> <li>• Promote “human in the loop” practices<sup>60,61</sup></li> </ul>
PHASE 4: Maintenance	<ul style="list-style-type: none"> <li>• Lack of algorithmic accountability<sup>47,62</sup></li> </ul>	<ul style="list-style-type: none"> <li>• Performance surveillance<sup>33,63,64</sup></li> <li>• Organization surveillance governance<sup>65</sup></li> </ul>

Figure 12: AI Risks and Mitigation Practices across the AI Implementation Cycle<sup>224</sup>

## II. AI Stacks

Another key aspect of AI governance discussion is AI stacks. AI stack is defined as a comprehensive combination of tools, libraries, and solutions used to develop applications with GenAI capabilities. AI stacks include programming languages, model providers, LLM frameworks, vector databases, and more.<sup>225</sup> For example, an AI stack can be the process where a general-purpose model, such as GPT-4, serves as the foundation for more specialized applications like hiring systems.<sup>226</sup>

In these cases, both the provider and the user of the AI stack are generally responsible for its operation. However, if a specific component within the stack fails, the provider of that component may also share responsibility.<sup>227</sup> Experts have emphasized the need for those building on general-purpose AI systems to seek detailed information and enforceable

guarantees regarding the system's performance for functions. One MIT policy brief highlights that regulatory and liability frameworks should aim to clarify situations where user responsibility is appropriate, especially when the AI system is used in unintended ways.<sup>228</sup> Providers are expected to specify proper uses, implement best-practice guardrails, and clearly define their legal responsibilities. Users must also be aware of the acceptable uses of AI systems.

### III. General-Purpose AI Systems

Concerns have arisen over general-purpose AI systems, like GPT-4, regarding their disclosure of potential uses and safeguards against unintended applications. Given the broad applicability and potential risks associated with general-purpose AI systems, such as chatbots with human-like interactions, experts suggest regulations might necessitate that these systems disclose their intended uses and implement safeguards to prevent unintended applications.<sup>229</sup> Providers may also be required to monitor their AI systems and report issues, akin to how pharmaceutical companies track their products. Specific concerns, such as realistic deep fakes and advanced surveillance, might require stricter regulations and clear labelling of AI-generated content to address risks that are distinct from those posed by human actors.<sup>230</sup>

### IV. AI Agency

Although an unpopular opinion, some experts have also discussed the possibility of the federal government establishing a new agency specifically for AI oversight.<sup>231</sup> Such an agency could have a narrowly defined scope to address the broad applicability and complexities of AI regulation, employing technical experts to advise existing regulatory bodies on AI issues. Alternatively, an existing agency with a relevant regulatory mission could be tasked with AI oversight, provided it is independent. Another option is a self-regulatory organization, like the Financial Industry Regulatory Authority (FINRA) in the financial sector, which could develop and enforce standards under federal supervision. In this case, regulation of AI systems for specific applications may continue to fall under existing agencies.

Currently, various organizations are preparing to mitigate AI-enhanced security threats: The MITRE Corporation launched an AI Assurance and Discover Lab<sup>232</sup> in March 2024 and collaborated with Microsoft<sup>233</sup> on the Adversarial Threat Landscape for AI systems (ATLAS)<sup>234</sup> in 2023. At this time, collaboration between MITRE and CISA or NIST has not been announced.

### V. Intellectual Property (IP)

Experts maintain that developing beneficial AI systems necessitates a clear framework for IP rights to ensure that human creativity remains incentivized.<sup>235</sup> Current legal IP standards



affirm that only humans can hold IP rights, meaning AI itself cannot own such rights. However, the application of existing IP laws to AI-generated content, especially regarding copyright, remains uncertain. AI has the potential to significantly increase instances of copyright infringement, complicating how creators can protect their work and identify potential infringements.<sup>236</sup> For example, Microsoft pledged to cover any copyright infringement claims against users of its GenAI products, provided users follow the established guardrails and content filters.<sup>237</sup> The AI EO attempts to mitigate this issue. It assigned the Under Secretary and Director of the United States Patent and Trademark Office to publishing a guide for patent examiners as well as to conducting a study that recommended steps to mitigate AI-related copyright issues.<sup>238</sup>

### C. Findings from Expert Interviews

Expert interviews with 10 cybersecurity and AI professionals supplemented the report’s findings from research. The expert interviews reveal the necessity of developing flexible and collaborative approaches to address biases in AI, emerging cybersecurity risks, and the rapid advancements in real-time AI content generation. Experts also emphasized the importance of balancing innovation with risk management, highlighting the role of public-private partnerships and adaptable guidelines. Furthermore, the interviews provided valuable perspectives on global regulatory approaches, the evolving nature of AI safety, and the need for practical, real-world governance frameworks. Table 17 summarizes these key themes from the expert interviews.

Table 17: Key Findings and Implications from Expert Interviews

Theme	Implication for Governance Approaches
Bias in Data <sup>xxxviii</sup>	Bias in data inputs and outputs can exacerbate barriers for marginalized individuals. Governance must ensure fairness, accuracy, and protection against these biases.
AI-Related Cyber Threats and Risks <sup>xxxix</sup>	AI poses challenges and solutions for cybersecurity, including threats related to social engineering, misinformation, election integrity, financial institutions, and supply chain vulnerabilities. Governance must address these risks comprehensively.
Advancements in Real-Time AI Generation <sup>xl</sup>	Significant advancements in real-time AI content creation are expected in the next few years, necessitating updated governance frameworks to manage new challenges in cybersecurity and content regulation.
Risk Management Frameworks vs. Prescriptive Regulations <sup>xli</sup>	A risk-based voluntary approach with flexible guidelines is preferred over rigid regulations, allowing for innovation while addressing high-risk applications. Public and stakeholder engagement can enhance effectiveness.

<sup>xxxviii</sup> Government Agency 1 and AI Policy Expert 1 interviews.

<sup>xxxix</sup> Government Agency 1 interview.

<sup>xl</sup> Government Agency 1 interview.

<sup>xli</sup> Government Agency 1 interview.

Governance Oversight <sup>xlii</sup>	Board and senior management oversight is key for AI governance, particularly in risk management and audit. Like the recent collaboration with the SEC on cyber requirements, addressing gaps at the board level is crucial for AI governance.
International Regulatory Differences <sup>xliii</sup>	The EU adopts a precautionary approach to protect citizens, exemplified by the General Data Protection Regulation (GDPR), and often introduces new regulations based on existing ones. In contrast, the UK favors a more liberal approach, focusing on regulation only when there is a clear risk to avoid stifling innovation. The US has fewer data protection laws to guide AI regulation. The challenge is finding the right balance of intervention and regulation to suit market needs and innovation.
Regulating AI vs. Product Safety <sup>xliv</sup>	AI regulation must address ethical considerations and sector-specific challenges beyond traditional product safety while adapting to technological advancements.
Risk-Based Regulation vs. Over-Regulation <sup>xlv</sup>	Focus on high-risk AI applications first to establish effective guardrails rather than trying to regulate all AI uses simultaneously. This approach, seen in the EU's shift back to a high-risk-first strategy, emphasizes assessing AI applications by their potential harm. Regulation should target use cases rather than the technology itself, adapting to emerging risks and changes through multi-stakeholder guidance.
Responsibility Across the AI Value Chain <sup>xlvi</sup>	All parties, including developers and users, share responsibility for AI systems. Developers should ensure their models are well-trained, monitored, and documented. Accountability should be embedded throughout the value chain, with the potential for regulatory approval or self-governance to foster trust.
International Collaboration and Multistakeholder Approaches <sup>xlvii</sup>	Effective AI regulation requires global cooperation and involvement from all stakeholders to develop standards that account for diverse ethical considerations and societal impacts. Regulatory sandboxes can help test and refine regulations in a controlled environment, fostering trust and enabling better regulatory decisions.
Lessons from Other Sectors <sup>xlviii</sup>	Sectors like financial services with experience in model regulation can offer insights for managing AI. Other sectors, such as agriculture, may lack this level of expertise.
US Departments in AI Regulation <sup>xlix</sup>	US technology-neutral laws allow agencies to regulate AI across sectors. Coordinating across agencies poses challenges due to competing regulators.
NIST's Role in AI Governance	NIST is actively working on AI governance by developing use cases for its RMF and establishing the US AISI for red teaming and benchmarking. These efforts aim to address technical and regulatory challenges, though the process is still developing. The collaboration with international frameworks and focus on public-private partnerships are crucial for advancing AI governance.
Drawing on Existing Cybersecurity Organizations and Regulatory Approaches <sup>l</sup>	AI model safety evaluation is still evolving, with early models predicting threats like early cybersecurity predictions. As AI and cybersecurity share foundational similarities but differ in maturity and complexity, AI may develop regulatory strategies akin to those in cybersecurity. Understanding AI threats and effective safety evaluations will require continuous adaptation as the field matures.

<sup>xlii</sup> Government Agency 1 interview.

<sup>xliii</sup> AI Policy Expert 2 and AI Policy Expert 1 interviews.

<sup>xliv</sup> AI Policy Expert 2 interview.

<sup>xlv</sup> AI Policy Expert 2 and Data Consultant 1 interviews.

<sup>xlvi</sup> AI Policy Expert 2 and Data Consultant 1 interviews.

<sup>xlvii</sup> AI Policy Expert 2, AI Policy Expert 4 and AI Policy Expert 1 interviews.

<sup>xlviii</sup> Government Agency 1 and Data Consultant 1 interviews.

<sup>xlix</sup> Data Consultant 1 interview.

<sup>l</sup> Security Expert1 and AI Policy Expert 5 interviews.

Mitigation of AI-Enabled Mis- and Disinformation <sup>li</sup>	Effective mitigation approaches include legislation targeting deepfakes specifically, investing in content provenance and watermarking, enhancing media detection, and improving digital literacy. A layered, adaptive approach is essential as AI evolves.
Vendor Responsibility and User Dependence	Human users, who do not have much experience and practice with evolving AI technologies, may be tempted to overly depend on AI tools. Governance approaches should include mitigation of risks created from overdependence. Vendors should play a role in misuse prevention and potential impact explanation.

## D. Approaches to Evaluating and Considering AI Governance Measures

Additionally, several organizations have attempted to provide criteria and evaluation approaches to AI Governance Measures, including MIT, the World Economic Forum (WEF), and the Centre for Security and Emerging Technology (CSET). This report reflects the critical take aways about AI governance approaches and adapts them to propose an AI governance criterion in the following sections.

### I. MIT: A Framework for US AI Governance

In November 2023, MIT released a policy brief titled “A Framework for US AI Governance: Creating a Safe and Thriving AI Sector.”<sup>239</sup> The policy brief was motivated by two key objectives: 1) to maintain US leadership in AI, and 2) to ensure the broad development of AI in ways that are beneficial across various domains. The policy brief argues that a combination of regulation and liability law is essential to ensure AI is developed and used in ways that promote its long-term benefits. The report outlines several guiding principles<sup>240</sup>:

- **Alignment with Existing Norms and Regulations:** AI governance should develop alongside AI technology by extending current regulatory frameworks to cover AI applications, rather than creating entirely new regulations, as seen in the EU.
- **Applying Current Regulations to AI:** Existing laws governing human activities should be extended to AI in relevant domains like healthcare and finance to ensure that AI systems are regulated like human actions while addressing risks and preventing circumvention of current laws.
- **Extending Legal Frameworks to Government AI Use:** Government activities—like policing or hiring—should adhere to extended legal frameworks when involving AI.
- **Enforcement by Existing Authorities:** AI regulations should be enforced by the same entities that govern human actions after the entities develop AI expertise.
- **Stricter Standards for AI Capabilities:** AI’s unique capabilities may need stricter regulations than those for humans, particularly in areas like pattern recognition.

<sup>li</sup> Election Analyst 1 and Policy Expert 2 interviews.

- **Disclosure of AI’s Intended Purpose:** AI providers must disclose the intended purpose of their systems before deployment, with guidance from regulatory agencies or through case law, to ensure transparency and accountability.
- **Defining AI for Regulation:** AI should be defined based on its functions (i.e., content generation) to determine which systems are subject to regulations.
- **Developing Auditing Regimes:** Auditing frameworks should be created to assess AI systems for issues like bias with standards set by appropriate entities like NIST.
- **Auditing System Development:** An auditing ecosystem might develop through mandatory audits by government or users, or organically through market demand and legal liability, ensuring intellectual property protection.
- **Prospective vs. Retrospective Audits:** Different types of audits, like prospective (pre-use) or retrospective (post-use), have distinct limitations and requirements, necessitating clear guidelines and accountability.
- **AI Interpretability Over Explainability:** While full explainability of AI decisions may not be possible, systems should be made more interpretable to provide insights into how outcomes are reached, with regulatory encouragement.
- **Training Data Quality:** The quality of training data is crucial, and AI systems should be designed to mitigate issues like bias and inaccuracies. Testing, monitoring, and auditing can help address problems stemming from flawed data sources.

## II. WEF: Presidio AI Framework

In January 2024, The WEF presented the Presidio AI Framework which provides a structured approach to the safe development, deployment, and use of GenAI. <sup>241</sup> The framework underscores the importance of shared responsibility of four key actors—AI model creators, adapters, users, and AI application users—for early identification of risks, proactive risk management, and timely implementation of effective guardrails. The Presidio Framework includes three core components<sup>242</sup>:

1. **Expanded AI Life Cycle:** Establishes a comprehensive view of the entire GenAI life cycle, highlighting the different actors and responsibilities at each stage.
2. **Expanded Risk Guardrails:** Outlines robust guardrails to be implemented throughout the AI life cycle, with a focus on prevention rather than mitigation.
3. **Shift-Left Methodology:** Advocates for applying guardrails at the earliest stages of the AI life cycle, adapting a software engineering concept to promote broader GenAI adoption.

The framework focuses on foundation models and integrates risk mitigation strategies across the entire AI life cycle, from creation and adaptation to eventual retirement. Grounded in extensive research on the AI landscape and informed by input from a diverse community of stakeholders and practitioners, the framework emphasizes the importance of established safety guidelines and recommendations, especially with technical details.

## Expanded AI Life Cycle

The expanded AI life cycle combines elements from data management, foundation model design and development, release access, the use of generative capabilities, and adaptation to specific use cases. Figure 13 below illustrates the Presidio AI Framework’s expanded AI life cycle, and each phase is detailed below<sup>243</sup>:

- **Data Management Phase:** Establishes the foundation for responsible AI ranging from data access to data type cataloguing to navigate laws in model creation.
- **Foundation Model Building Phase:** The model progresses through stages from design to internal audit, each with specific guardrails.
- **Foundation Model Release Phase:** Focuses on responsible dissemination and risk mitigation, classifying models by access levels, from fully closed to fully open. Each access level has distinct norms, standards, and challenges.
- **Model Adaptation Phase:** This phase describes stages, techniques, and guardrails for adapting a pre-trained foundation model to perform specific generative tasks.
- **Model Integration Phase:** Involves integrating the adapted model with an application and developing APIs downstream.
- **Model Use Phase:** Users interact with hosted models via natural language prompts, emphasizing the need for robust guardrails established during earlier phases while adapters may add further safeguards based on specific use cases.

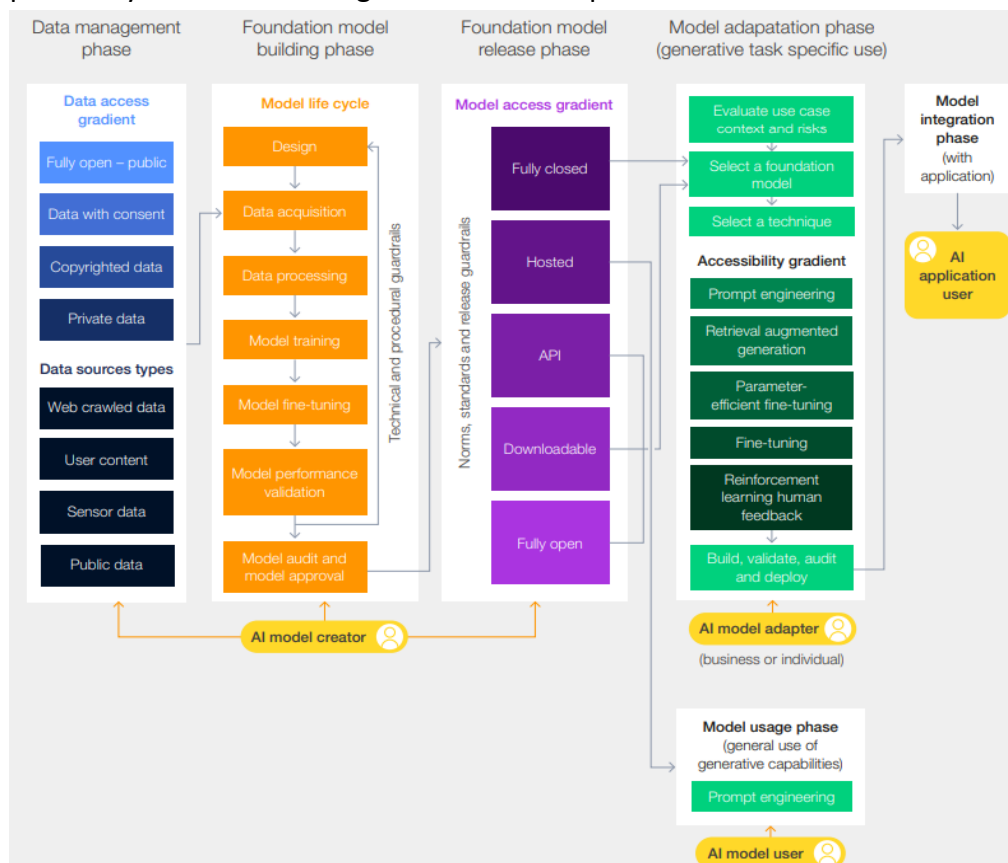


Figure 13: Presidio AI Framework’s expanded AI Life Cycle<sup>244</sup>

Unexpected model behavior at any phase can harm users and lead to reputational or legal consequences for both the user and the model creator or adapter. The likelihood of misuse—such as plagiarism, intentional non-disclosure, IP violations, deepfakes, generation of malicious content, and misinformation—grows as model access shifts from fully closed to fully open if vigilant oversight is not maintained. The Presidio Framework outlines some of the safety benefits and challenges of different model release types, as seen in Figure 14.

Release type	Safety benefits	Safety challenges
<b>Fully closed</b>	Creators control the model use and can provide safeguards for data privacy and the IP contained in the model. There is more clarity around responsibility and ownership.	Other actors have limited visibility into the model design and development process. Auditability and contributors' diversity are limited. Application users have minimal influence on model outputs.
<b>Hosted</b>	Creators can provide safeguards for model outputs, such as blocking model response for sensitive queries. They can streamline user support. Use can be tracked and used to improve model responses.	Similar challenges as "fully closed". Other actors have little insight into the model, limiting their ability to understand its decisions.
<b>API</b>	Creators retain control over the model while empowering users to adapt the model for specific use cases. They can provide user support. This level of access increases the "researchability" of the model. Increased access allows users to help identify risks and vulnerabilities.	Even though transparency is limited, model details can be inferred by third-party tools or attacks (in case of bad actors).
<b>Downloadable</b>	Along with creators, adapters and users are also empowered through the release of model components. This means more transparency, flexibility for model use and modification of the model.	Lowered barriers for misuse and potential bypassing of guardrails. Model creators have difficulties in tracking and monitoring model use. Users typically have less support when experiencing unexpected undesirable model outputs/outcomes.
<b>Fully open</b>	These models provide the highest levels of auditability and transparency. This level of access increases global participation and contribution to innovation – also in terms of safety and guardrails. Adapters and users are empowered to adapt models that better align with their specific task and improve existing model functionality and safety via fine tuning.	These models present a higher chance of possible misuse. Access to model weights means higher risk of model replication for unintended purposes by bad actors. Ambiguity around accountability and ownership.

Figure 14: Safety Benefits and Challenges of Model Release Types<sup>245</sup>

### Guardrails Across the Expanded AI Life Cycle

The framework's second component provides technical and procedural guardrails for distinct phases of the AI life cycle, emphasizing that a combination of both types is needed to ensure safe systems. Technical guardrails ensure the technical quality and consistency of AI systems, whereas procedural guardrails maintain process consistency and control. The framework provides some examples of guardrails and their placement, shown in Figure 15.

Highlighted guardrails	Phase placement
Red teaming and reinforcement learning from human feedback (RLHF) <sup>3</sup>	Building
Transparent documentation and use restriction	Release
Model drift monitoring and watermarking	Adaptation

Figure 15: Select Guardrails and their Phase Placement<sup>246</sup>

### Model Building Phase

In the model-building phase, the framework asserts that early red teaming is crucial to ensure model safety and address vulnerabilities, including prompt injection and toxic content. This approach addresses vulnerabilities and ethical concerns early in the AI lifecycle, building trust among stakeholders. For foundation models, tests should include prompt injection, data leakage, jailbreaking, hallucinations, and toxic content identification. Although red teaming effectively addresses known vulnerabilities, it may not identify unknown risks before mass release.<sup>247</sup> Notably, the NIST RMF refers to various aspects of this model building phase, highlighting risks from information security, intellectual property management, bias in outputs, and more. The NIST SSDF also included recommendations for protecting AI-related resources and data from AI-specific attacks such as data poisoning.

Incorporating RLHF early in the process provides a strategic advantage by facilitating efficient learning and faster iterations, which improves model performance and alignment with human objectives.<sup>248</sup> RLHF involves training a reward model to fine-tune the primary model, resulting in more desirable responses and a reliable iterative feedback loop involving human raters, a trained reward model, and the foundation model. While RLHF enhances performance, it risks introducing new biases and raises data privacy and security concerns.

Novel methods for implementing these guardrails include “red teaming language models with language models” and reinforcement learning from AI feedback (RLAIF), which use language models to generate test cases or provide safety feedback.<sup>249</sup> These techniques automate the process, reducing the time needed for implementation, and can be applied in later phases as well. Using them early allows for adjustments to model hyperparameters, though they may introduce new, unidentified vulnerabilities.

### Model Release Phase

In the model release phase, guardrails include protective measures for downstream actors. Transparent documentation involves detailing decisions, processes, and data related to the AI model. This transparency allows downstream users to understand the model’s limitations, assess its impact, and make informed decisions. Best practices include developing persona-

based templates, gathering information throughout the life cycle, and using tools like datasheets, data cards, and model cards to improve documentation and auditing.<sup>250</sup> Automation can enhance efficiency, though challenges include determining the most relevant information and balancing proprietary versus required disclosures.<sup>251</sup>

Use restriction focuses on limiting the model's application to prevent misuse and unintended harm, such as harmful content generation and inappropriate model adaptation. Effective practices include implementing restrictive licenses (e.g. responsible AI licenses), tracking model use, providing clear usage guidelines, and incorporating feedback and incident reporting mechanisms.<sup>252</sup> Additionally, the framework maintains that moderation tools should be used to filter or flag undesirable content, prevent harmful prompts, and block misaligned responses.<sup>253</sup> Challenges in this area involve developing comprehensive licensing standards and high-quality tools to manage model responses.<sup>254</sup>

### Model Adaptation Phase

In the model adaption phase, a key objective is to ensure that the modified model continues to be effective and aligned with its intended use case. Model drift monitoring is essential for maintaining performance as it involves regularly comparing post-deployment metrics to address issues such as evolving data, adversarial inputs, and noise.<sup>255</sup> Best practices include employing data, algorithms, and tools to track data drift, as well as defining protocols and adaptation techniques to manage model performance and uphold customer trust.<sup>256</sup> Watermarking of model outputs is another important consideration, with its application depending on factors like the use case, model type, and watermarking objectives. Watermarking embeds hidden patterns to help detect and mitigate the mass production of misleading content.<sup>257</sup> It assists in identifying AI-generated content for policy enforcement, attribution, legal action, and deterrence. However, workarounds like removing watermarks or paraphrasing can undermine its effectiveness. Thus, while watermarking can be implemented earlier in the model creation phase for ownership purposes and adapted later to control visibility, various layers of AI risk management approaches must be implemented throughout the AI model lifecycle.

### **Shifting Left for Optimized Risk Mitigation**

The WEF applies the shift-left concept to GenAI models and expands with the following<sup>258</sup>:

- **Rising Interest in Foundation Models:** As foundation models become prevalent, model creators and adapters can be different entities, requiring early-stage safeguards.
- **Increased Model Accessibility:** With powerful models accessible to users with varying technical skills, there is a greater demand for transparency and clear documentation.
- **Elevated Risks:** Users face risks from using factually incorrect outputs without validation, potential misuse in disinformation campaigns and adversarial attacks (e.g., jailbreak).



For GenAI, the shift-left approach suggests implementing guardrails earlier in the life cycle to mitigate risks at each phase. Depending on the model’s purpose, there may be a trade-off between guardrails and safety. Figure 16 shows shift left steps for GenAI model creation:

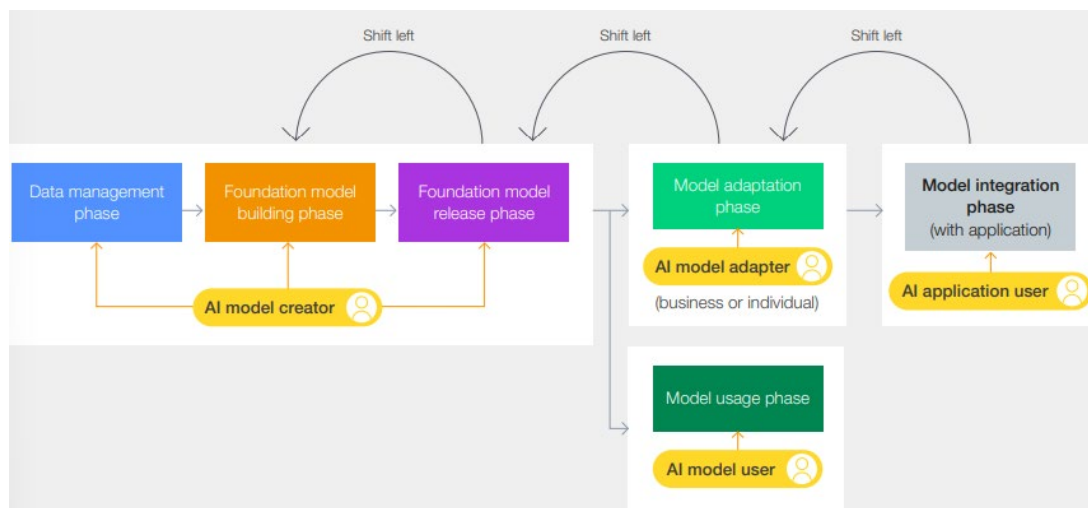


Figure 16: Presidio AI Framework with shift-left methodology for generative AI models<sup>259</sup>

The shift-left methodology, demonstrated by the arrows arching to the left, proposed in the Presidio framework includes three key steps<sup>260</sup>:

- **Release to Build Shift:** The AI model creator proactively incorporates guardrails throughout the foundation-building phase and collects necessary data and models facts and transparency surrounding these instead of beginning these mitigation methods in the foundation model release phase.
- **Adaptation/Use to Release Shift:** The AI model creator, instead of the adapter, incorporates additional guardrails, establishes norms and standards for use, and creates comprehensive documentation to help downstream actors understand and make informed decisions regarding model use during the foundation model release phase.
- **Application to Adaptation Shift:** The AI model adapter, instead of the AI application user, proactively incorporates guardrails considering the use case and considering the documentation from AI model creators about the foundation model During the model adaptation and usage phase.

### III. CSET: Report on Flexible Approach

In July 2024, CSET released a report advocating for the US to take a flexible approach to AI governance. The report offers three principles for US policymakers to follow<sup>261</sup>:

- **Know the terrain of AI risk and harm:** Use incident tracking and horizon scanning from industry, academia, and government to assess AI risks and harms, and gather data to

inform governance and manage risks effectively.

- **Prepare humans to capitalize on AI:** Educate policymakers and the public on AI opportunities, risks, and harms to ensure responsible and lawful use of AI applications.
- **Preserve adaptability and agility:** Create adaptable policies that can evolve with AI advancements, avoiding overly restrictive regulations and preventing regulatory capture that could stifle innovation and competition.

### **Know the Terrain of AI Risk & Harm**

The first principle to “know the terrain of AI risk and harm” outlines a few approaches:

1. Capture data on AI harms through incident reporting.
2. Invest in evaluation and measurement methods to strengthen our understanding of cutting-edge AI systems.
3. Build a robust horizon scanning capability to monitor new and emerging AI developments, both domestically and internationally.

#### Capture Data on AI Harms through Incident Reporting

The report asserts that regulators should prioritize the collection of data on AI incidents to inform policy and ensure innovation isn't stifled. This comprehensive approach involves incident reporting, evaluation science, and intelligence collection. AI systems should undergo rigorous testing to identify potential issues such as drift or malicious misuse. The authors conceptualize AI harm on a spectrum of minimal to existential risks and provide four categories that lawmakers can use in AI governance<sup>262</sup>:

1. Demonstrated harms
2. Probable harms involving known risks in deployed AI systems
3. Implied harms where studies could uncover new weaknesses
4. Speculative harms, including existential risks

Incident reporting would involve gathering data from AI-related accidents and harms through mandatory, voluntary, and citizen reporting. A public incident reporting system would exclude military and intelligence AI incidents, which would have separate, secure reporting channels. Federal agencies like the Securities and Exchange Commission (SEC) could oversee mandatory and voluntary reporting while citizen reports might be collected through government systems or NGOs.

#### Invest in Evaluation & Measurement Methods

The authors of the report note that evaluating AI systems is still in its early stages; however, they advocate for public, private, and academic investment in basic research to advance this field and develop standardized methods and toolkits for AI developers and regulators. Understanding the trustworthiness properties of AI systems, such as robustness, fairness, and

security, is crucial for policymakers to create effective governance mechanisms.<sup>263</sup> The establishment of the US AISI is a promising development, but it may currently lack the resources needed to fully meet its objectives as outlined in EO 14110 and related policy guidance.<sup>264</sup>

#### Build a Robust Horizon Scanning Capability to Monitor New Developments Globally

The report also suggests effective horizon scanning can aid US legislators and regulators in adapting to new risks and potential harms by providing early insights into emerging AI technologies and trends. A key component of this strategy is the establishment of an open-source technical monitoring center.<sup>265</sup> This center would support the US intelligence community and other federal agencies by tracking AI progress across commercial, academic, and government sectors. It would facilitate better integration of open-source and classified information, enhancing overall intelligence gathering and interpretation.<sup>266</sup> For intelligence agencies, the focus might be on technologies that impact military systems, while other agencies might monitor AI applications with significant implications for economic competitiveness and societal well-being.<sup>267</sup> By staying informed on new capabilities, US policymakers can more effectively respond to emerging challenges, particularly in competition with nations like China or other states with potentially harmful AI applications.

#### **Prepare Humans to Capitalize on AI**

The second principle for AI governance aims to improve use and education of AI and has two objectives—to develop AI literacy among policy among policymakers and the public.

First, developing AI literacy among policymakers is crucial to have a foundational understanding of various AI models, their strengths, limitations, and potential biases. Policymakers should be aware of how AI systems can fail unexpectedly and the challenges of transparency and explainability. This knowledge will help them identify suitable uses for AI and understand how AI inputs can influence human decision-making.<sup>268</sup> Training and curricula designed for policymakers can enhance their AI literacy skills and prepare future leaders to address emerging regulatory challenges.<sup>269</sup>

Similarly, the public's AI literacy is equally important for ensuring responsible interaction with AI systems. Educating citizens from an early age and continuing through adulthood will help them understand AI's potential and limitations, especially in fields where AI is increasingly applied. Awareness of when to trust AI outputs or remain skeptical is essential as is familiarity with risks like plagiarism or copyright issues. Looking at successful public AI literacy programs, such as those in Finland, can provide valuable insights.<sup>270</sup> Additionally, educating the public about the dangers of AI-generated disinformation can serve as a defense against its misuse, protecting democratic processes and societal integrity. AI developers should be mindful of the risks related to integrating models into products.<sup>271</sup>

## **Preserve Adaptability & Agility**

The report's third principle asserts that policymakers must adapt and incorporate new knowledge into governance efforts, allowing the flexibility to iteratively update policies as technology evolves.<sup>272</sup> Thus, the authors outline three key recommendations:

1. Consider where existing processes and authorities can already help govern AI if certain implementation gaps are addressed.
2. Remain open to future AI capabilities that may evolve in new and unexpected ways.
3. Consider the costs and tradeoffs involved when planning AI governance approaches.

### Consider Where Existing Processes & Authorities Can Help Govern AI

AI governance may require new regulations, but existing laws can often be adapted for AI-related issues, offering advantages in speed and familiarity for lawmakers.<sup>273</sup> For instance, current regulations under bodies like the FTC or the US Food and Drug Administration may already address some AI concerns, such as copyright infringement or discrimination. Policymakers should identify gaps in existing legal structures and assess where new resources or procedural changes are needed to effectively manage AI's impact.<sup>274</sup>

When existing frameworks are insufficient, developing flexible and adaptive regulations is crucial. The report notes that NIST's AI RMF serves as an example of an adaptable policy.<sup>275</sup> Additionally, leveraging state and federal regulations to gather data and try experimental governance approaches can help address emerging AI challenges effectively.<sup>276</sup>

### Remain Open to Future AI Capabilities that May Evolve Unexpectedly

Policymakers should stay adaptable to the evolving landscape of AI capabilities as future advancements may not follow the same trends as past developments. Recent progress in LLMs has largely been driven by algorithmic improvements and significant increases in computing power, which has been costly. However, future advancements might emerge from novel algorithmic innovations or improvements in data processing that require less computational power. As the growth in computing usage for training large models appears to be slowing, policymakers should remain attuned to emerging trends through sources like open-source collection, incident reporting, and horizon scanning to ensure effective regulation of new AI developments.<sup>277</sup>

### Consider the Costs and Tradeoffs in Planning AI Governance Approaches

When designing AI governance frameworks, lawmakers need to weigh the costs and tradeoffs of different approaches. Accurate estimation of the labor and resources required is crucial to selecting a feasible strategy. One key concern is regulatory capture where agencies designed to serve the public interest end up favoring the commercial interests of the industries they regulate. This bias can lead to policies that benefit the regulated entities rather than the public.

While input and cooperation from AI companies are valuable for identifying trends and risks, lawmakers must be cautious of potential biases these companies may bring. Regulatory capture is a significant risk, and avoiding it requires maintaining a large, skilled government workforce for tasks such as risk assessment and testing.<sup>278</sup> This effort can also be challenging and costly. Alternatively, shifting responsibility for testing and risk mitigation to firms could reduce government costs but may lead to standards that favor large companies. Lawmakers must balance these considerations and decide if a high-intensity, government-focused approach is needed to mitigate risks or if a less stringent approach is sufficient.<sup>279</sup>

## E. Criteria for AI Governance Approaches

AI governance aims to mitigate the risks associated with AI technologies while ensuring innovation and maintaining fairness, transparency, and accountability. AI risks vary significantly across applications, leading to conflicting priorities. For example, addressing socioeconomic risks like bias and discrimination may require different strategies than tackling cybersecurity threats such as data breaches or misinformation. Therefore, the effectiveness of governance frameworks depends on the specific use of AI, necessitating a hybrid and phased approach that accommodates different use cases, outcomes, and sectoral differences.

Findings from the investigation into US AI regulations, research of AI framework, and interviews informed the development of the following criteria for AI governance approaches.

Three pillars are essential—foundational principles, higher-level strategies, and sectoral use case-based—in the order listed. Figure 16 outlines a proposed framework for developing AI governance approaches and regulations. Expert interviews and research findings indicate that all three are needed to form effective and practical approaches to AI governance. An optimal approach will combine principles-based guidance with prescriptive measures, including regulation tailored to specific use cases and industries, as this detailed focus has been less explored in current frameworks.



Figure 17: Pillars to Develop Effective AI Governance

First, all regulations should adhere to the foundational principles such as multi-stakeholder involvement, explainability, and adaptability, including mechanisms for periodic risk reassessment. Second, higher-level strategies should focus on robustness, international harmony, and comprehensive risk management, areas where the US has already made progress. Finally, sector-specific prescriptive regulations are crucial for addressing particular use cases and outcomes, an area where the US will need to invest further, drawing parallels with the approach used in cybersecurity.

## I. Pillar 1: Foundational Principles

Pillar 1 seeks to establish foundational principles that all AI regulations follow to ensure a consistent and equitable governance approach. This pillar includes three key components:

- **Multi-Stakeholder Involvement:** Engage stakeholders—civil society, industry, and policy experts—to contribute to the development and oversight of AI regulations.
- **Explainability:** Mandate that AI systems operate transparently, providing clear explanations for their decisions and actions.
- **Adaptability:** Implement mechanisms for updates to regulations with periodic risk assessments and adjustments based on evolving AI technologies and emerging risks. AI governance should be built on established frameworks related to privacy and security rather than treating AI as new. In looking forward, AI technologies will continue to change, and AI foundational principles should be able to adapt to apply to evolved AI capabilities. For instance, in the US, AI governance has been built upon adaptable foundational principles such as the NIST Cybersecurity Framework.

## II. Pillar 2: Higher-Level Strategies

Pillar 2 provides components to guide the development of broad strategies to address key risks and ensure international alignment in AI governance. This pillar has three components:

- **Robustness:** Create strategies aimed at minimizing all types of AI-related risks, including bias, cybersecurity threats, and other operational vulnerabilities.
- **International Harmony:** Align with global standards and collaborate with international bodies to create harmonized frameworks across jurisdictions, avoiding fragmentation and enabling companies to effectively navigate global regulations.
- **Comprehensive Risk Management:** Prioritize AI risk management across the AI life cycle, emphasizing human centricity, social responsibility, and sustainability.

### III. Pillar 3: Sector-Specific Prescriptive Regulations

Pillar 3 seeks to develop detailed, use case-specific regulations tailored to different AI applications and sectors. This Pillar contains three components:

- **Sectoral Focus:** Implement regulations that address the unique risks and requirements of specific industries, such as healthcare, finance, and transportation.
- **Use and Outcome-Based Guidelines:** Create prescriptive rules based on the particular use cases and outcomes of applications of AI within each sector, ensuring that regulations are relevant and effective for different scenarios. A use case and outcome-based approach tailors AI regulations to the specific impacts of each application, rather than relying on broad, uniform rules.
- **Long-Term Development:** Recognize that developing these sector-specific regulations will take time, drawing from approaches used in other fields like cybersecurity to build a robust regulatory environment.

## F. Assessment of Current US AI Governance Measures

While one way to assess AI governance approaches is to map the coverage of AI-enabled risks to the measures—as done in Section 6.C’s Table 16—given the layered and hybrid US approach, assessment may be against the criteria proposed in Section 7.E. The following section assesses four US AI governance approaches—AI EO, AIBoR, NIST RMF 1.0, and NIST SSDF—with the proposed three-pillared approach derived from the research and interview findings detailed in Sections 7.B through 7.D.

The analysis is presented on a color-coded table. Green indicates strength in the pillar. Yellow indicates that the pillar’s components are included in the governance approach but with gaps such as weak implementation, lack of specificity, exclusion of applications. Red indicates weakness in the pillar or the lack of the pillar’s components.

The AI EO demonstrates strength in Pillar 1: Foundational Principles. However, the wide breadth of the EO makes it a weak high-level strategy or sectoral use case-based approach.

Table 18: Assessing the AI EO Against Proposed Governance Criteria

Pillar 1: Foundational Principles	Pillar 2: Higher-Level Strategies	Pillar 3: Sectoral Use Cases
<b>Multistakeholder Engagement:</b> integrated feedback from civil society and companies.	<b>Robustness:</b> defines AI systems broadly including GenAI and neural networks, focuses on harms overall.	<b>Sectoral Focus:</b> distributes entities across government entities which may be sector-focused, i.e., DOE.
<b>Explainability:</b> directs 50 federal entities with more than 150 requirements.	<b>International Harmony:</b> developed a US vision in AI governance, signaling to the international community the US approach.	<b>Use and Outcome-based:</b> none.
<b>Adaptability:</b> focuses on safety, security, privacy, equity, civil rights, workforce development, competition, and responsible government use.	<b>Comprehensive Risk Management:</b> Focuses on resulting rights and risks more than the development phase or other phases of the AI lifecycle.	<b>Long-term Development:</b> none.

The AIBoR also demonstrates strength in Pillar 1. In terms of its strength as a higher-level strategy, the AIBoR supports comprehensive risk management but fails to cover a robust set



of risks by excluding industrial and operational applications of AI. In terms of its strength as a sectoral use case-based guidance, the Bill of Rights does include sector-specific guidance and some use and outcome-based guidance. Finally, there are no explicit indications of the AIBoR being updated according to a set timeline.

Table 19: Assessing AIBoR Against Proposed Governance Criteria

Pillar 1: Foundational Principles	Pillar 2: Higher-Level Strategies	Pillar 3: Sectoral Use Cases
<b>Multistakeholder Engagement:</b> collaboration between the OSTP, academics, human rights groups, the general public and even large companies like Microsoft and Google. <sup>280</sup>	<b>Robustness:</b> excludes many industrial and/or operational applications of AI	<b>Sectoral Focus:</b> includes sector-specific guidance with a focus on most high-priority algorithmic harms across healthcare, financial services, education and housing.
<b>Explainability:</b> clarifies that the AIBoR should only apply to automated systems that have the potential to meaningfully impact the American public’s rights, opportunities, or access to critical resources or services, expands on specific examples as well.	<b>International Harmony:</b> USAID launched an AI Action Plan to embed risk mitigation in AI and support responsible technology worldwide, provides a foundation for future AI regulation internationally.	<b>Use and Outcome-based:</b> coordinated across the government to publish inventories of non-classified government AI use cases to adhere with civil rights and privacy laws (OMB, OSTP, Federal Chief Information Officers Council).
<b>Adaptability:</b> emphasizes safety, fairness, privacy, civil liberties and rights, covers a wide range of issues and federal actions that can be leveraged to expand capacity in the future and provides a foundation for future AI regulation in the US and internationally	<b>Comprehensive Risk Management:</b> covers algorithmic discrimination, data to mitigating risk at the hiring level and process.	<b>Long-term Development:</b> no specific plans for update.

The next two tables evaluate the NIST RMF 1.0 and NIST SSDF. Overall, the NIST RMF appears to be a strong foundational document, higher-level strategy, and sectoral use case-based guidance. The NIST RMF demonstrated strength in eight of the nine components and only showed weakness in long-term development; however, NIST tends to publish follow ups and updates to their documents and guidelines.

The NIST SSDF is not the strongest governance approach<sup>liii</sup> to AI risks, likely because it focuses on AI and software development. The NIST SSDF does incorporate multistakeholder engagement and is appropriately explained. However, it is not as robust or applicable as other guidance due to its foundation in software development practices. The NISF SSDF does not include sector-focused nor use and outcome-based guidance. Despite the lack of inclusion of the proposed nine pillars, the NIST SSDF is still an appropriate and unique approach to mitigating AI risks.

Table 20: Assessing NIST RMF 1.0 Against Proposed Governance Criteria

Pillar 1: Foundational Principles	Pillar 2: Higher-Level Strategies	Pillar 3: Sectoral Use Cases
<b>Multistakeholder Engagement:</b> developed through collaborative approach with workshops and input opportunities.	<b>Robustness:</b> flexible with relevance across diverse use cases	<b>Sectoral Focus:</b> companion document (NIST AI 600-1 AI RMF) includes cross-sectoral profile.
<b>Explainability:</b> developed through transparent process with Request for Information, drafts for public comments, and workshops	<b>International Harmony:</b> resonates with frameworks form EU, Singapore, and OECD.	<b>Use and Outcome-based:</b> companion document (NIST AI 600-1 AI RMF) includes use-case profiles by function and category, tailored to requirements, risk tolerance, resources.
<b>Adaptability:</b> includes issues such as mitigating hallucinations, data privacy, environmental impact, information integrity, and more; builds on previous knowledge on privacy and security.	<b>Comprehensive Risk Management:</b> covers the different stages of AI lifecycle as well as business risks.	<b>Long-term Development:</b> NIST tends to publish follow ups and updates to their documents.

<sup>liii</sup> Another framework like the NIST SSDF that is not focused on AI but mentions AI-related components, is the WEF’s Data Equity Framework. This Framework emphasizes the need to identify and mitigate bias throughout the data life cycle from AI model testing to training. The Framework suggests various methods including embedding model and system traceability and accountability and disclosing non-human interactions as good ways to mitigate bias. The WEF specifically proposes that AI models and data processes be audited with clear documentation to ensure that the analytical process can be reviewed—this approach aligns with the NIST’s focus on auditability and traceability of AI systems. The document is a good high-level guidance to start discussions and guide strategic decisions on AI model testing but lacks technical details such as model performance metrics, robustness of adversarial attacks, and specific testing protocols.

Table 21: Assessing NIST SSDF<sup>281</sup> Against Proposed Governance Criteria

Pillar 1: Foundational Principles	Pillar 2: Higher-Level Strategies	Pillar 3: Sectoral Use Cases
<p><b>Multistakeholder Engagement:</b> NIST typically engages various stakeholders for its publications, and the EO required NIST to solicit private sector, academia, and public sector input.</p>	<p><b>Robustness:</b> covers various risks at the software development level, aimed at software producers and acquirers only.</p>	<p><b>Sectoral Focus:</b> none.</p>
<p><b>Explainability:</b> aligned with business mission requirements, organizational goals, risk tolerance, and available resources; helps identify gaps and guide a prioritized action plan.</p>	<p><b>International Harmony:</b> refers to ISO documents.</p>	<p><b>Use and Outcome-based:</b> only defines risks at a high-level, not use case focused.</p>
<p><b>Adaptability:</b> based on software development practices (BSA, OWASP, SAFECode)</p>	<p><b>Comprehensive Risk Management:</b> focuses on software security but spans the AI model development process (data sourcing to training to software integration); does not include AI deployment or operation</p>	<p><b>Long-term Development:</b> NIST tends to publish follow ups and updates to their documents.</p>

## 8. Conclusion

Building on the phase 1 report on AI-enabled cyber risks and through this investigation of AI-enabled risks to elections, the two reports highlight that AI will exacerbate existing risks from disruptive attacks to information operations. In the US, the various governance approaches have led to potential confusion and inefficiencies; yet, as shown in Table 16, the multi-layered approach has enabled mitigation measures against the AI-enabled risks. At a high-level, the US approach covers all five AI-enabled cyber risks to a certain degree:

- AI-enhanced traditional cyberattacks are covered by preexisting cybersecurity regulations as well as the AI EO's focus on ensuring security of AI technology.
- The three bills that have passed the markup stage in the Senate Committee on Rules and Administration—Protection Elections from Deceptive AI Act, AI Transparency in Elections Act, and the Preparing Election Administration for AI Act—are aimed at regulating deepfakes and AI-enabled disinformation. Furthermore, the AIBoR promotes the ethical use of GenAI which is an attempt to prevent malicious actors from spreading AI-enabled disinformation.
- In terms of AI-enabled disruptions and maloperations of systems, the AI EO is extensive on mitigating inherent biases of AI models and systems and the AIBoR focuses on safe and effective systems.
- AI-enabled national security threats from military applications of AI, AI-enabled terrorism, AI-enabled bioterrorism is accounted for by the NIST AI RMF as well as the NIST Plan for Global Engagement on AI Standards that can encourage international allies to implement a standard for AI.
- NIST SSDF for GenAI focusing on secure development practices for GenAI and dual-use foundation models can help mitigate against vulnerable code generation and dissemination as well as other AI-enabled business risks.

However, this report's in-depth research into and interviews of US AI governance measures reveal that the currently existing AI governance approaches are better categorized into three pillars: 1) foundational principles, 2) higher-level strategies, 3) sector-specific prescriptive regulations. The report emphasizes the importance of having all three pillars for a voluntary-base, multi-layered approach informed by various stakeholders. Foundational principles should be an explainable and adaptable baseline for strategies and sector-specific regulation to build upon. Higher-level strategies should be robust, pursue international harmony, and include comprehensive risk management. Finally, sector-specific prescriptive regulation should address the unique risks and requirements of a sector, be use and outcome-based guidelines, and be developed in the long-term.

Based on this categorization, this report proposes a criterion for effective AI governance and framework measures. In assessing a select number of the US AI governance measures

according to the proposed criteria, the report concludes that the NIST RMF is a strong documentation of foundational principles, higher-level strategies, as well as use-case and sectoral-based guidance.

Furthermore, all of the US AI governance approaches demonstrated strengths in including public-private partnership and multistakeholder engagement. The private sector, such as OpenAI and Anthropic, is in formal collaboration with US AISI and NIST regarding AI safety research.<sup>282</sup> Beyond this level of coordination and the private sector's general compliance with US government guidelines, additional details are not typically publicized.

To further strengthen AI governance approaches, the public private partnership should focus on the following:

- Research Coordination: Public and private stakeholders should actively work to design incentive structures that facilitate greater coordination between academic researchers and the private sector throughout the technology development lifecycle.
- Support Open Innovation and Transparent Knowledge Sharing: Policymakers and AI providers should contribute to frameworks to democratize AI through responsible sharing of resources, including data, source code, models, use cases, and research findings. Also both the public and private sectors should encourage the sharing certification processes, ensuring transparency and trust among stakeholders.

At this time, the US AI governance approaches all face an implementation challenge as well as lacks a binding enforcement mechanism. The AI EO is allegedly fully implemented, but researchers were only able to find evidence of around 80% implementation. The AIBoR calls upon various federal agencies to address AI-related issues, but only some agencies have adequately responded. As mentioned, the Department of Labor limited its focus to surveillance related to labor organizing, neglecting broader employee surveillance concerns. Similarly, the AI EO, AIBoR, NIST RMF, and NIST SSDF are nonbinding principles that request voluntary adoption which can limit the impact and effectiveness of the governance approaches. There needs to be a mechanism for monitoring and evaluation that measures the implementation progress and evaluate impacts of governance measures. This monitoring can help make adjustments as necessary to address emerging challenges and opportunities.

Finally, as AI technologies continue to evolve, long-term development of AI governance approaches need to be incorporated to ensure that the governance evolves with the technology.

## 9. Expert Interviews

### AI Policy Expert 1

AI Policy Expert 1 is Lead in the Data Policy team (Centre for the Fourth Industrial Revolution) and Lead of the "Resilient Governance and Regulation" working group of the AI Governance Alliance. The expert is passionate about exploring the ways in which AI, data and technology in general can promote digital transformation of governments, cross-sector collaboration, user-centered services, and inclusive development around the world. Prior to joining the Forum, The expert worked at the Inter-American Development Bank (IDB), supporting digital government and statistical capacity building projects. Before that, she worked at the World Bank, OECD, Ashoka Changemakers, and the Mexican Ministry of Foreign Affairs. The expert holds a master's degree in International Affairs, and bachelor's degrees in political science and in International Relations.

**We are beginning to see AI-related framework and governance measures such as the Biden EO and the NIST AI Risk Management Framework. What are the strengths and weaknesses of US AI regulations and frameworks?**

I think in the absence of more binding overarching regulation in the US, the executive order and the work that NIST has been doing are sending a message in the right direction. It's saying, 'We want to address this and ensure we develop technology responsibly.' It's also about bringing all key players into the conversation to make it more of a dialogue. I think there's an emphasis on responsible AI development across all stages of the AI development cycle, which sends the right message. There's a focus on protecting individual rights and privacy. When GDPR came out, Europe was perceived as more rights-based, whereas the US was seen as more market-based, favoring companies. With the EO and what NIST is doing, there's an understanding that we need to assess risk and issues of responsibility and protection to safeguard individuals. This, in my opinion, is changing the perception.

The approach the US is taking aims to balance the promotion of innovation—empowering companies to experiment and develop solutions—while emphasizing responsibility, accountability, and transparency. It's an important balance, though achieving perfect alignment in practice may be challenging. One inherent weakness is the absence of a comprehensive national AI policy in the US. It's complex with federal and state levels having their own regulations, but there's a need for alignment and harmonization to cover citizens across the entire territory.

Another aspect that countries, including the US, should aim for is promoting multinational collaboration to align risk metrics and standards with other nations. This involves navigating different national AI priorities, security issues, data localization, and governance approaches. It's not an easy task, but striving for alignment with other countries, finding minimal common ground, and fostering interoperability are crucial, considering data, technology, and talent flow across borders with varying political ideologies. Regulations should also consider the practical realization that perfect alignment may never be achievable.

**Considering the challenges of harmonizing regulations across the US and globally, what would you suggest as the factors that must be included in a comprehensive AI approach?**

I want to emphasize that when I talk about alignment, I understand that there won't be a single regulatory approach or a universally accepted set of principles, especially not around AI. It's more about fostering collaboration across countries and advancing that type of alignment. However, there are limitations and inherent fragmentation due to specific national interests. Thus, collaboration across standard-setting organizations is crucial. In the absence of a specific national AI strategy, countries could still collaborate to develop softer regulatory frameworks and align with others. Understanding the needs of the private sector and other actors, and how these translate into cross-border issues, is also key. Consulting with different sectors and actors, understanding specific use cases and the unique regulatory challenges they pose, is essential. Recognizing self-regulation or self-governance innovations in the country, and how they align with international practices, is relevant.

It's not about making exceptions or making things easier just because we don't want to lower standards. We want to ensure that technology is developed responsibly by establishing clear regulations, codes, and guidance that provide incentives for actors to comply. It also involves providing training, building capacity, allocating funding, and supporting open research. Access to resources such as computing power and data is essential to facilitate the development of these technologies.

Therefore, any conditions or incentives that the public sector can create to empower different actors to comply with regulations and develop technology responsibly will ultimately benefit everyone. This approach ensures that companies can effectively meet regulatory requirements and respond to the demands of responsible technology development.

**What are factors that hinder AI governance frameworks currently?**

Technology moves so fast, and regulators are often steps behind. Keeping regulators updated on technological advancements is crucial. It involves capacity building and transparency from companies in their development processes. Flexible regulatory mechanisms are essential due to ongoing technological progress. New solutions could include self-governance with government collaboration, and regulatory sandboxes for testing regulations in controlled environments. This approach avoids delays in advancing regulations, aligning with agile methodologies. National AI policies are crucial anchors alongside soft guidance and best practices. Maintaining a constant dialogue with diverse stakeholders is challenging but vital.

**Could you just explain how a regular sandbox would work, especially in the context of AI?**

Although I'm not an expert in the topic or directly involved in development, I can provide an example. Regulatory sandboxes allow companies and governments to collaborate in testing specific regulations and product development. It's an iterative process where they assess feasibility within project cycles, ensuring compatibility and clarifying requirements. Countries like Singapore, UAE, Brazil, and the UK have pioneered these sandboxes.

### **How effective are AI safety practices and measures by NIST and CISA in suppressing AI misuse?**

Concerning NIST, the organization has been proactive in developing guidelines and frameworks that extend beyond the US. For example, I recall a recent discussion with a government official who leads AI initiatives where they highlighted collaborative efforts with NIST in Singapore. Together, they've crafted robust risk management frameworks. There's also been alignment noted with the UK in similar endeavors. They incorporate crucial elements such as fairness, accountability, transparency, trustworthiness, and security, which are essential. They're setting a standard for the responsibility that should be upheld. In this regard, I believe they're effectively conveying this message.

### **Can you expand on some of the work you do around the socioeconomic factors of AI, and how AI governance approaches seek to regulate those impacts?**

At the AI Governance Alliance, we collaborate extensively with our community, which comprises over 300 members. In the working group I lead focusing on regulation and governance, we have about 130 members. Our approach aims to incorporate diverse perspectives on a global scale, emphasizing best practices and regulatory recommendations that extend beyond just economic or social impacts of AI. We strive to consider how regulations affect entire societies and various stakeholder groups while also exploring ways to make regulations more agile. I would say our work not only covers the issues you mentioned but goes a step further. However, given the rapid evolution of the topic, it can be challenging to track every specific development at the country level. Instead, we focus on understanding broader trends and approaches to AI governance, aiming to strike a balance between providing overarching insights and actionable recommendations applicable across different countries.

### **How can AI models be governed or tested to make sure they consider all these impacts?**

That's a tricky one. Safety and security are key, so there should be regular audits and privacy assessments. Assessing consent, ensuring effective initial consent is respected, and addressing traceability issues regarding data integrity and mutation within systems are important for privacy protection. Another challenging aspect is testing for biases and ensuring fairness, especially given that many models are inherently biased due to historical data biases. We attempt to correct these biases through adjustments in models, assessments, and weight corrections, although these assessments are often conducted at the end of the data lifecycle. Ideally, such assessments should occur throughout the entire AI development cycle, integrating perspectives beyond just engineers and developers to encompass diverse team viewpoints. In practice, overcompensation for bias can lead to models that do not accurately reflect historical facts. Incorporating sociologists, psychologists, and bias training for team members is essential, alongside implementing procedures for continuous bias assessment throughout processes. Establishing partnerships among stakeholder groups for consultation and testing is crucial, facilitating iterative learning and system refinement. Transparency, accountability, and traceability in addressing these issues are paramount, along with establishing clear, measurable criteria for assessing and addressing safety and security. Continuous monitoring should not be merely corrective but preventive, leveraging user feedback and model outcomes. Open communication with stakeholders is essential for accountability—both internal and external processes and reporting to governmental bodies on issues like data breaches and other concerns are crucial.



## **Do you think there is sufficient support and effort directed towards establishing criteria for the entire lifecycle of data?**

I think there's an intent. I don't know if there are enough resources devoted to it—perhaps due to aggressive product launch timelines and pressure to respond to political correctness or issues. I think there could be a lot more efficiency if there was an exchange of these practices, more transparency about how companies conduct these assessments and where they stand, and the metrics they're measuring and exchanging those experiences. But I feel like there are still a lot of barriers to open conversation for many reasons, including business secrets, intellectual property, and the risk of giving away sensitive information to the competition. So, in that sense, maybe academia, society, and government could help build those bridges, have more alignment, and promote more effective and aligned measurements that can help test models for privacy, security, safety, biases, inclusion, all of that. An active strategy involves assessing these companies to allocate financial resources for designing metrics, running them, and collecting data to assess impact. Collaboration with other sectors in different countries should focus on developing metrics that are more relevant for the models throughout the AI development cycle, rather than just corrective measures at the end.

### **Security Expert 1**

Security Expert 1 is Chief Executive Officer of a security consulting firm, a global cyber security advisory, training, consulting, and media services company supporting hundreds of major organizations across the world. The expert recently retired from a telecommunication company after thirty-one years of service, culminating as Senior Vice President and Chief Security Officer from 2004 to 2016. The expert was elected a Fellow in 2010 and is a Research Professor in the Computer Science Department at the NYU Tandon School of Engineering, and a Senior Advisor at the Applied Physics Laboratory at Johns Hopkins University. The expert is the author of six books on cyber security, and dozens of major research and technical papers in peer-reviewed journals and conference proceedings. The expert has also been Adjunct Professor of Computer Science at the Stevens Institute of Technology for the past twenty-nine years, where he has introduced over three thousand graduate students to the topic of information security. The expert holds the BS degree in Physics, the MS/PhD degrees in Computer Science and is a graduate of the Columbia Business School. He holds ten patents in cyber security technology, and he served previously on the Board of Directors for a bank and the NSA Advisory Board. The expert's work has been highlighted on CNN, the New York Times, and the Wall Street Journal. The expert has worked directly with four Presidential administrations on issues related to national security, critical infrastructure protection, and cyber policy.

## **How should AI model safety evaluation be approached or tested?**

We don't know yet what the AI threats really are. There are some preliminary situations that hackers, researchers, and academics are trying to get a sense of through model projections. Early models predict that AI threats would include data source pollution, making algorithms come to the wrong conclusion, and hallucination. This is similar to the early days of cybersecurity: we had predictive models for risks, and some predictions were correct while others were not. We are going to have to wait and see what the threats are to really identify the approach to AI model safety evaluation.

**What are the key differences between traditional cybersecurity approaches and the regulatory needs emerging with AI technologies?**

The biggest difference is maturity. There is a lot in common between the two fields in how they originated. When cybersecurity became an issue, people initially tried to find similarities with standards, commercial products, and best practices, but we learned that cybersecurity needed a different approach. I believe that after these early days of AI, we may go down the same route as well especially in terms of safety and security. Another difference is that AI is dealing with algorithms and techniques that we don't understand as much as cybersecurity. For example, what really happens when we fire neurons in a neural network. This means that forensics will be different as well.

**What do you think are the weaknesses of current AI frameworks and governance measures?**

The current ones are too academic right now. For example, there is a presumption of what a machine learning operation pipeline looks like in business, but if you asked a business about that pipeline, you will get a lot of different answers. The idea that there is a standard way to build models is just not true, and that is what MITRE, NIST, and the AI EO are missing. Also, day-to-day practitioners can't change what Google, for example, is doing. Practitioners just must use the output. The best that a framework can do is to work with technology providers to ensure that the output is somewhat clean. In the early days of cybersecurity, there was the orange book—it was a framework for cybersecurity that eventually became defunct and completely thrown out. We must be willing to take such steps with AI framework when they don't work. We must be flexible and open to change.

**What changes would you make to current AI governance measures?**

Right now, most of the governance measures are principles. What is missing is practical experience. Companies that use AI daily need to be in tune with the challenges and use cases—both the good and bad use case. Most of the time, AI is not necessarily bad. It's just that there is bad security. AI can help expose risks, but then we need a framework to deal with this as well as guardrails. Jumping too quickly to regulation and legislation may stifle innovation, especially when there is so much potential for healthy AI use. To have people use AI for good use cases, we need to avoid overregulation, be intentional about international agreements, and make key societal decisions with the good use cases in mind.

**A majority of the experts are highlighting the importance of having use cases. What are some efforts on this front?**

Professor Fei-Fei Li at Stanford invented ImageNet which trains databases and machine learning to match good training examples or use cases. There needs to be a common language for some kind of use cases repository. For example, having a score of 1 to 10 with 10 being a terrible outcome. We also need to avoid over concluding from use cases and using that conclusion to dictate policy. The next 10 years will be a lot of balancing the good and bad use cases, and the cybersecurity community needs to allow for mistakes to happen instead of blocking everything.

## **What are some AI-enhanced election risks?**

There are a few different dimensions or layers to this. There is the base layer which includes the infrastructure that moves votes and reports to election servers. The infrastructure is dependent on networks, systems, cloud, computers, and software. A thousand things can go wrong there, but we did a good job in 2020 under the leadership of Chris Krebs to keep the base layer secure. Then, there is the fake news later that can influence votes without dealing with the election infrastructure. Then, there is the actual voting that uses software to tally votes, keep a voter database, and more. We are still using old technology and processes for these layers. The one factor that keeps us safe is that the US has distributed elections. If we had a centralized election, one collapse or risk can breakdown the whole election. With our 50 different elections, problems don't cascade from one to another.

## **AI Policy Expert 2**

AI Policy Expert 2 is the Policy lead for Artificial Intelligence and Machine Learning - AI Governance Alliance at an international organization. Prior to her role at the organization, the expert held the position of Head of EU Corporate Affairs at a software company from 2022 to 2024. From 2020 to 2022, the expert was the Head of Policy and Operations for Technology and Advanced Manufacturing at the Department for International Trade (DIT) in the United Kingdom. Earlier, The expert managed the Telecoms and Trade EU Exit strategies at the Department for Digital Culture, Media, and Sport (DCMS). The expert's policy journey began as a Senior Policy Adviser at the Broadband Stakeholder Group, followed by the expert's role as a European Affairs Executive at the Architects Registration Board. With a robust background in guiding cross-sectoral policy frameworks and a deep commitment to advancing the responsible integration of AI technologies into global markets, the expert continues to be a leading voice in AI governance and policy innovation.

## **Can you tell us about your background and current work at the WEF?**

I joined the forum about two months ago. My background is in policy development and strategy, and I've been working on behalf of a multinational organization that specializes in developing AI applications across various sectors. As part of that work, I was deeply involved in the negotiations of the EU AI Act. Currently, I serve as the policy lead on AI within the AI Governance Alliance at the World Economic Forum. I'm managing two distinct projects. The first focuses on regulatory governance and aims to establish best practices across different regulatory frameworks, providing guidance for policymakers and industry leaders. The second project is broader and centers on what we refer to as "Inclusive AI for Growth and Development." This initiative explores how countries in the Global South can develop national AI strategies while integrating effective governance mechanisms.

## **Based on your work at the AI Governance Alliance, how does your work seek to balance innovation and regulation in its approach to shaping AI policies?**

That's a big question, but I want to start with a caveat. The World Economic Forum's role in this context is not to determine the best regulatory framework for AI or to advise policymakers on a single approach. Instead, we examine the various existing regulatory frameworks for AI. Our focus is not just on the pros and cons of each but on understanding why so many different approaches exist and in

what contexts they were developed. We also explore potential areas for convergence on regulatory issues at the global level. I can't definitively say which regulatory approach is the best. However, I can outline the different levels of regulatory intervention. For instance, we have principle-based approaches, rule-based frameworks like the EU AI Act, and other horizontal regulations from different nations. Then, there's the UK model, which deliberately chose not to regulate AI across all sectors. Other countries have opted out of regulating AI altogether, fearing it could stifle innovation.

Addressing your question about the impact of regulation on innovation is complex. It really depends on the regional context and the underlying regulations that apply to various uses of AI. Taking the EU as an example, it is currently the most interventionist in terms of AI regulation, being the first attempt at a comprehensive horizontal framework. This includes establishing rules for AI system development, implementation, and risk assessment. Initially, the EU's approach was risk-based, but it has become more ambitious over time. Industries are concerned that such regulation might hinder innovation. While that concern is valid, we need to consider the purpose of regulation in this context.

In the EU, regulators adopt a precautionary approach to protect citizens, reflecting a broader cultural attitude towards regulation. This is evident in various digital regulations, with GDPR being the most prominent. Existing regulations often justify the introduction of new ones despite challenges. In contrast, the UK government made a conscious choice not to impose extensive regulations on AI to avoid negatively impacting innovation. The UK's regulatory culture is more liberal, focusing on intervention only when there is a clear risk, unlike the EU's precautionary stance. In the U.S., the context is different again, with relatively few data protection regulations that could serve as a foundation for AI regulation. However, what we have observed over the past two or three years is that AI regulation is now a priority for governments worldwide. They are all seeking to develop appropriate guardrails. The challenge remains: how much intervention is necessary and what level of regulation is suitable for the specific market context, the ways AI is used, and its potential impact on innovation?

**What are some of the specific challenges associated with rule-based and principle-based frameworks, as well as any other approaches you've encountered?**

There are several challenges associated with regulating AI, and from my perspective, the most significant one is the inherent uncertainty surrounding the technology. Unlike product safety regulations, where we can often predict worst-case scenarios, assessing the risks of AI is far more complex. Many risks remain unknown and difficult to anticipate.

One challenge regulators face is whether to regulate AI in the same manner as product safety or to adopt a different approach altogether. It's essential to recognize that AI regulation extends beyond product standards; it intersects with ethics and various aspects of life and different sectors. This complicates the regulatory landscape significantly. For example, we need to consider how data is collected, used, and processed, as well as the specific risks associated with AI models themselves. Can these models become autonomous? Can they behave in ways we cannot foresee? These questions highlight the need for tailored regulatory responses.

The biggest challenge, therefore, is the sheer number of unknowns related to AI. Creating regulations that can endure over time is virtually impossible given the rapid evolution of technology and its applications across various industries and societal contexts. Regulators must continuously monitor emerging risks and advancements, which requires a substantial amount of information and analysis to inform effective decision-making. Furthermore, there are challenges related to the policy goals of regulations. While many regulators focus on AI safety as the primary reason for regulation, others, like the EU, consider the implications of AI on fundamental rights. This divergence in focus creates additional complexities. In summary, the key challenge is that AI is an evolving technology with many unknown risks attached, making effective regulation a difficult task.

**In the past two years, how has the understanding of generative AI's risks and benefits evolved, particularly regarding the need for democratization and regulation?**

I would say there have been various developments in the last two years. AI has existed for 30 years and was already regulated, primarily in areas like data protection and finance. However, generative AI became publicly accessible, which sparked panic regarding its potential effects. This was the primary experience regulators faced. Of course, generative AI has numerous applications that are extremely beneficial for industries and small businesses. Nevertheless, the fear and uncertainty about how generative AI could impact lives and society have dominated the conversation, leading to a rush toward regulation. I believe that if we take a step back and consider generative AI more carefully, we need to slow down and weigh the risks to the public alongside the benefits. It's essential to ensure that our regulations do not unintentionally limit applications that could be extremely useful for economic growth and societal development.

**What is your view on the need for regulatory frameworks or guidelines for users of AI?**

I think you need to establish some guardrails when it comes to AI regulation. The most effective approach is to begin by identifying the highest risks to society, focusing on potential harm and the dangers of specific uses of AI. By prioritizing the regulation of high-risk applications first, we can create a solid foundation for determining which AI systems should be prohibited and which ones require scrutiny from regulators.

What I meant by "panic" is the tendency to try to regulate all possible uses of AI simultaneously, regardless of their risk levels. This mindset was evident during the EU's negotiations on the AI Act, where there was a prevailing notion that AI was inherently unsafe because its applications were still largely unknown. However, industry representatives highlighted the many beneficial uses of AI across various sectors. This ultimately led the EU to revert to a high-risk-first approach, emphasizing the need to regulate the most concerning AI applications first. Broadly speaking, regulating specific technologies may not be the most effective strategy. Instead, we should start with use cases. By analyzing how AI will be applied across different sectors, we can better identify high-risk and low-risk situations. Low-risk AI systems, which pose little threat, generally do not require regulation.

It's important to avoid regulating AI simply because it's a new technology. Technology evolves rapidly, and in the near future, we might be discussing entirely different innovations. Thus, regulation should focus on how systems will be used rather than on the technology itself. Although we can't predict all

risks, incorporating mechanisms to assess and anticipate them within the regulatory framework is crucial. This should be an ongoing exercise for all regulators, and I believe this reflects both the EU's and the UK's approach: forming multi-stakeholder groups to guide regulators in identifying emerging risks. This allows for ongoing reclassification of risks, adapting to changes and different uses.

**You've stated that it makes sense to regulate use cases instead of technology as a whole—do you foresee a need to regulate both providers of harmful AI models and users who misuse AI?**

I believe everyone along the value chain has a responsibility. For instance, generative AI developers have a responsibility to ensure that their AI systems are properly trained, monitored throughout their lifecycle, and that all development processes are documented. It's easy to assume that once an AI model is deployed, the developer is no longer responsible for its usage—placing that burden solely on the user, but I don't think that's correct. Developers, particularly those working with generative AI, must ensure that their models are both developed and used appropriately. They also have a responsibility to communicate these guidelines to their users, which include not just individuals but also businesses that fine-tune or create new versions of the models. This principle of accountability should permeate across the entire value chain and is crucial to embed in our practices. There are various ways to enforce this accountability. For example, developers could be required to follow specific steps and obtain regulatory approval before bringing an AI system to market. Alternatively, we could encourage self-governance, urging companies to adopt transparency and accountability principles to foster trust in the absence of stringent regulations. Ultimately, the primary responsibility lies with the developers.

**Could you elaborate on how international collaborations or frameworks can enhance public-private cooperation and partnerships in this context?**

When discussing international collaboration or cooperation, we automatically and rightly think about the development of AI standards. This effort requires the involvement of the entire ecosystem—industry, civil society, and all stakeholders—because it's no longer just about product safety; it's about understanding how AI systems impact society. Different regions may have varying ethical considerations, which must be considered when creating these standards. Additionally, regulators need a foresight function to anticipate the implications of AI. In establishing a foresight function—whether through the EU AI office or another organization—it's essential to have full participation from all sectors of society. Industry plays a pivotal role here; when developing models, companies must understand their potential uses. Their insights are vital for informing policymakers about effective regulations. Regarding public-private partnerships in AI, regulatory sandboxes come to mind as a way to test different approaches before formal regulation. This requires the public sector to be fully engaged, allowing companies to pilot AI models under controlled conditions. This approach helps build trust with the private sector, demonstrating that regulation doesn't have to happen immediately and that there is room to evaluate products under the right conditions.

### **Can you expand on the private-public partnerships in AI regulation?**

When considering the EU framework or even the UK, they have established expert working groups. These groups consist of experts in AI development from various sectors, including specific companies, trade associations, civil society organizations, and consumer organizations. Their goal is to replicate societal dynamics and create long-standing expert groups that continuously inform regulators.

At the WEF, we aim to do something similar. We have over 208 organizations globally, with about 350 to 360 members in our community. We strive to be as representative and diverse as possible, not just within the AI ecosystem but also from a global perspective. This is challenging, yet it has led to many positive outcomes, such as fostering innovative solutions and addressing problems more swiftly.

I believe that as AI regulation evolves, we will increasingly see this multi-stakeholder approach, where governments rely on diverse input to inform their regulatory decisions. This method is not only faster but also more reliable than having the government make decisions without testing them in real-world scenarios.

### **Can you discuss how government regulators select stakeholders from the AI value chain to ensure effective collaboration and innovation?**

I believe it's essential for governments to carefully choose the sectors they engage with. Typically, you'd examine the value chain across AI, focusing on the major generative AI developers—of which there are only a handful. You always need to have them, or you need to consult them, because they are the ones driving innovation. Next, you have large players who may not develop AI models themselves but are fine-tuning them. Additionally, it's important to include businesses that use AI in both B2B and B2C contexts. Ultimately, it's the responsibility of the government and regulators to ensure they select the right stakeholders for a diverse group. Sometimes they succeed in this, but at other times, they miss the mark. This has certainly been the case in the EU, where it has been hit or miss. However, I believe that everyone is improving in this regard.

## Policy Expert 1

Policy Expert 1 currently serves as Director of Information Integrity for an ICT solution vendor's Democracy Forward Program. The expert is a former non-resident policy fellow with the Stanford Internet Observatory. The expert served as Senior Cybersecurity Advisor at the Department of Homeland Security, where the expert focused on election security issues. The expert previously served as a Commissioner at the Election Assistance Commission, including serving as the Commission's Chairman. Prior to that, the expert held staff positions with the Ohio Secretary of State's office, where the expert oversaw voting-system certification efforts and helped develop an online voter registration system. The expert holds a law degree and BS and BA degrees.

### **How have election-related threats evolved from 2016 Cambridge Analytica Era to the 2020 Twitter era to the 2024 generative AI era?**

The threat environment changed drastically. I led the election security work within the Department of Homeland Security (DHS) starting in 2018 but had already been working on the issue in 2016 as well. In 2016, the experience we had in the US was one in which Russia specifically targeted both campaigns and election infrastructure—the websites and voter registration databases of election authorities. They stole data, and in some cases, used that information for an information operation with troll farms and bots that were employed to spread false information about candidates and the election process. The tactic they used involved building their own audiences with fake accounts to disseminate misinformation independently. This operation was orchestrated by the Russian government as detailed in reports from the US Senate Intelligence Committee. They also organized or had Americans organize competing rallies on divisive issues such as racial politics, where opposing groups would converge to generate anger and frustration. As election time approached, the Russians shifted focus to political and electoral matters, aiming to bring the generated emotions and attention into the real world through news articles, media coverage, and actions by citizens.

Moving to 2018, my experience indicated a decrease in such direct engagement. Instead, within the US, social divisions and political polarization worsened, primarily manifesting as heated debates on platforms like Twitter. Similar tactics involving bots were employed, but the emphasis shifted to amplifying existing domestic controversies rather than creating new ones.

By 2020, the tactics had changed significantly. While some cyber activities persisted, the primary challenge in the information environment during the election was the widespread dissemination of claims about election fraud, ballot manipulation, and other electoral irregularities. These claims were amplified by both large domestic accounts and foreign state actors, exacerbating divisions within the country. For example, in Arizona, controversy arose over the use of Sharpie pens on ballots. A simple query on social media about the pens bleeding through ballots was exploited to sow doubt and suspicion about the election process. This tactic was not limited to Arizona but spread to other states, creating a narrative that Sharpies were being used to invalidate ballots and rig the election. Throughout 2020, narratives often emerged organically from smaller accounts and gained momentum until they were seized upon by larger influencers who validated and amplified them. This bottom-up approach to narrative development made it challenging to respond effectively, akin to a game of whack-a-mole for social media companies and election officials trying to moderate false information.



Ultimately, the sheer volume and speed of these narratives overwhelmed the information space, making it impossible for election officials to address every claim simultaneously. This pollution of the information space capitalized on existing online discussions rather than executing a pre-planned narrative by adversaries, demonstrating the evolving nature of information warfare tactics.

**How do you think the recent assassination attempt of former President Trump will affect AI-enabled influence operations?**

If we look at what we've observed with AI-enabled influence operations, there's been extensive concern about AI's impact on election information environments, prompting significant efforts by tech companies, nonprofits, and governments to understand and mitigate risks.

From what we've seen in recent elections like those in the EU, UK, India, and France, AI has been used in a couple of ways. First, there have been attempts to generate images, videos, and in some instances, manipulate audio, but the impact of these efforts has been limited. For example, during the Taiwanese election, the Chinese government attempted to use AI-generated content to influence discussions around election-related issues and candidates. Despite these attempts, interaction with such content was minimal and did not significantly influence the electorate. Similarly, in the EU and UK elections, AI was used sparingly and locally, primarily to provoke reactions or emotions rather than to deceive through deepfakes or falsified actions attributed to individuals.

In the United States, there were isolated incidents such as the use of AI-generated audio imitating President Biden in New Hampshire, suggesting election-related misinformation. However, these attempts were swiftly debunked and had little to no impact. Thus far, the widespread concerns about AI-driven deepfakes or deceptive AI have not materialized on a large scale. Instead, adversaries have utilized AI to amplify existing narratives and gain attention or emotional responses, albeit without overwhelming impact in most cases. In India, for instance, AI was observed in private chat groups but did not significantly alter the election discourse as anticipated.

Moving to the question of misinformation around the assassination attempt, there's been widespread speculation and conspiracy theories across the political spectrum. AI has played a minor role, mainly in creating exaggerated imagery. The greater concern remains mis/disinformation spread through traditional manipulative tools and platforms, like image editors. Therefore, whether or not it's AI-manipulated content, the strategies to mitigate their impact remain consistent. It is crucial to proactively flood the information space with trusted and authoritative sources, such as election officials' communications. For example, during the EU elections, public campaigns directed voters to verified sources like the EU Commission's website for reliable election information.

Learning from experiences in 2020, the key lesson is the continuous dissemination of trusted information well before and during elections. This proactive approach not only reduces the likelihood of misinformation going viral but also encourages critical scrutiny from the public when confronting dubious claims. In conclusion, whether addressing AI or traditional media manipulation, maintaining information integrity through proactive measures is paramount.

**What are the most effective methods for mitigating AI-enabled disinformation and information operations?**

We approach this similarly to cybersecurity. There's no one solution to tackle AI and information environment challenges. Instead, it requires a layered approach. First, we focus on providing indicators of trust to customers and voters regarding information. This involves directing them to trusted election official websites or reliable news outlets. We've heavily invested in content provenance—labelling media to indicate its origin, ensuring transparency in its creation and distribution. This approach allows media and voters to question authenticity if provenance is absent, which is crucial for enhancing trust. Additionally, we invest in watermarking for hidden metadata, which strengthens our verification capabilities. This hidden metadata serves as a digital fingerprint, enabling us to verify the authenticity of images and videos, even if the visible content is altered.

The second layer of our approach focuses on improving the detection of manipulated media. While this remains a significant challenge due to evolving AI technologies, we continuously enhance our detection capabilities. Our goal is to differentiate between trustworthy information and potential manipulations, providing clear labels to help voters navigate the digital landscape effectively. Lastly, we emphasize media and information literacy through comprehensive education campaigns. These initiatives empower voters, including demographics less familiar with technical aspects of media sourcing, such as older populations. By promoting critical thinking and responsible sharing practices, we aim to equip voters with the skills needed to assess online content accurately. This proactive approach is essential in safeguarding the integrity of elections and combating misinformation.

**Do you think that our society is doing enough to invest in digital literacy?**

I think we could always do more, so no. We need to understand that in media and information literacy space, the responsibility can't just fall on tech companies or civil society that don't have enough support. We are working with OpenAI and the Coalition on Content Provenance, and we've also collaborated with other civil society groups to fund their media and information literacy efforts. But there's also a role for government to play in this. We know that there's been consideration of adopting curriculum for media and information literacy. Some countries have done this effectively. I think government, civil society, and tech companies can work together to build an understanding within the population of how to interact with information on social media and on the web. Greater investment is needed, but it's not just money; it's the investment of time, educational resources, and government prioritizing this for their population. This is a long-term effort in media and information literacy. You can't just do it for a year and hope that it takes hold, it requires sustained effort and understanding.

**We have heard about the liar's dividend often in the context of AI-enabled influence operations. Could you speak about that and how you view the risk?**

This topic is quite critical, especially given recent events. During the 2020 US election cycle, we saw a troubling trend where many voters felt overwhelmed and unsure about what information to trust. This sentiment was pervasive at various stages of the election process, reflecting a broader challenge in our information ecosystem. When we think about the impact of AI on these issues, it's clear that AI has the potential to worsen these existing challenges. Mis/disinformation can spread more rapidly

and convincingly through AI-generated content. The real risk here is the erosion of trust among the electorate, leading to apathy and disengagement from civic discourse. To address these challenges, it's crucial to empower individuals with reliable sources of information. This means not only combating false narratives but also providing clear, transparent indicators of trust. Our commitment to enhancing content provenance is central to this effort. By offering verifiable information about the origin and authenticity of content, we aim to equip people with the tools they need to navigate today's media complexities. At its core, this is about ensuring that our democratic processes remain robust and resilient. When people have access to trustworthy information, they can engage meaningfully in debates, make informed decisions, and hold elected officials accountable. This foundational trust in information sources is essential for the healthy functioning of our democracy, safeguarding against the divisive impacts of misinformation and fostering a more informed electorate.

**What do you think is more likely to be a risk or threat for the US: foreign threat actors (Russia, China, and Iran) or domestic influence operations similar to those seen in the EU elections?**

It's hard to say because they are so intertwined at this point. We know that our foreign adversaries use our domestic divides to their advantage. I read an article today about the assassination attempts on former President Trump and how the Russians jumped to amplify circulating conspiracies. From 2016 to 2020, they realized they don't have to generate a lot of their own content; they can just amplify what's already out there. So, they go hand in hand. We're going to continue to see a pattern where our adversaries identify that content and play both sides against each other, as they have over the years, to create division and distrust. Their clear purpose is to portray American democracy as weak and to use this to message to their own population: 'You don't want this; you don't want democracy. It's not healthy. Look how it's playing out in the United States.

### AI Policy Expert 3

AI Policy Expert 3 is the policy research manager at a university's Institute for Human-Centered Artificial Intelligence (HAI), where the expert develops and oversees policy research initiatives. The expert is passionate about harnessing AI governance research to inform policies that ensure the safe and responsible development of AI around the world—with a focus on research on the privacy implications of AI development, the implementation challenges of AI regulation, and the governance of large-scale AI models. Prior to joining HAI, the expert worked as a China-focused consultant and analyst, managing and delivering in-depth research and strategic advice regarding China's development and regulation of emerging technologies including AI. The expert holds a Master's in International Policy and a Bachelor's in Chinese Studies.

**Your piece "Transparency of AI EO Implementation: An Assessment 90 Days In" emphasizes the importance of implementation and transparency. Can you speak on why these two factors are so critical and how they can help AI governance more effectively?**

When Biden's EO on AI came out, it was a massive document with lots of different tasks assigned to different agencies. We wanted to make sense of that in a somewhat methodical way and be able to track how the government is implementing it. Previously, HAI and Stanford's law school had done work on previous AI-related legal documents: two EOs and one AI in government act. Retroactively,

HAI tracked the implementation of the three measures and found inconsistent and poor implementation. Additionally, finding public evidence of implementation was difficult. Reporting implementation allows external stakeholders to provide informed advice to the government. Furthermore, weak and inconsistent implementation implies that leadership was not sufficiently empowered or lacked the resources to implement the regulations. This previous work showed us the importance of tracking implementation, if at least to pressure the government to make improvements.

In the last eight months since the Biden EO was put out, we are seeing a better and proactive implementation approach and progress reporting. The White House is putting out regular fact sheets and proactively talking about what it is doing; agencies set up websites to show what they are doing on the EO. Most importantly, things are being implemented. A variety of road maps are being created. Major policies, task forces, AI safety boards are being created. Clearly the full range of tasks are in motion. We are generally positive about this momentum, and we can see the AI EO is a government priority. Agencies are empowered, have resources, and are pressured to meet the deadlines. Government is better at reporting. One weakness is that we still cannot find all the information on implementation completion even though the US government claims 100% implementation. HAI was only able to confirm 70-80% of task completion.

**In examining the implementation of the Biden EO on AI, which tasks did you find most impactful and effective?**

In focusing on implementation, there is risk in making the process a box ticking exercise. There is a pressure to meet the deadlines from the EO but some activities were already underway or, for example, there is a requirement for each agency to assign a Chief AI Officer—for the most part, agencies are doing this, but perhaps, the new Chief AI Officer was the Chief Information Officer beforehand. It is hard to assess, but some significant activities include OMB's new policy on the use of AI in government which came out a couple months ago. It is a large and comprehensive effort. NIST came out with some documents. Various agencies like DHS and HHS now have roadmaps that feel meaningful and substantive. It is important to note, however, that the early requirements of the EO were foundational tasks: set up task forces, boards, and working groups or get stakeholder input. This groundwork lays the foundation for more meaningful work down the road.

**How are agencies working together, aligning, and/or collaborating for AI regulation, especially in the context of the Biden AI EO?**

Generally, I am not sure if there is a strong mechanism in place for this process and collaboration, but the predominant approach is to avoid making new agencies/departments/units. There is a question across the globe regarding whether AI should be dealt with under new regulatory agencies or if the existing organizations should adapt to cover AI. The EO approach is that AI work sits with every single part of government and that everyone has to work and input from their perspective. The teams and offices likely do have the right people who can take on the new AI tasks. The EO is also prescriptive at points telling which agency to work on what or with whom. The main mode for pushing forward this work has been to create task forces—not a new entity like department but identifying point people and empowering them to work on cross-cutting work. There is not much public information on this

yet, but the cross agency board of all Chief AI Officers will be an interesting mechanism for collaboration. However, it is too early for an external party to say how that is functioning so far.

**Are there any other criteria needed for measuring the effectiveness of AI governance/framework besides implementation and transparency?**

One factor to focus on is the AI talent within the government. Any AI governance framework is only effective and possible for implementation if you have the right expertise in government to implement and drive it forward. AI talent is a real issue all around the world. All these governance frameworks with hundreds of requirements and deadlines and thoughtful risk frameworks can exist, but if we don't have the technical or other talent to be able to implement or constantly iterate on it, you are going to just run into roadblocks, especially because AI technology evolves so quickly. This EO was great, comprehensive, and bold, but the government can only do it if there is talent. This means we need to look at immigration policies, internal policies for how you get people into government and what the pipeline is, rotate people in and out of government, or be able to pay people more to compete with the private sector. The EO was more foundational, so it didn't need the specific AI talent yet. Now, the government is reaching out to external technical advisors for input which is okay about public sector talent is highly necessary.

**What are the immediate next steps to further the effectiveness of AI governance measures?**

Ultimately the EO is not an actual legal framework, meaning it does not regulate. The EO mobilizes the federal government to think about how to tackle AI, create roadmaps, and think about policies. The clear next step is what concrete regulation will come out of this. Will there be binding regulations and requirements that the private sector must abide by? All we have thus far are voluntary frameworks and commitments. In the US, we are now going into political and unstable times where not much is going to happen in the election months or even before. But there is a dynamic range of bills being proposed by various senators that are all sitting there waiting to be moved through later. Ultimately, the EO built capacity in the government as a wakeup call to force agencies to consider how AI relates to their work? The EO emphasizes that departments need to build up teams/task forces and publish reports/guidance, but the interesting next step is what concrete regulatory tools will be created that will put requirements on companies and developers.

One thing to note is the tension between what is technically feasible and what we can realistically do. Specific implementations that are more technical are ongoing in the technical parts of the government like NIST and NTIA. NTIA looks at specifically how to govern open foundation models compared to closed models. NTIA has already engaged in stakeholder consultation on this and came out with a report.

## AI Policy Expert 4

AI Policy Expert 4 is a research fellow at a university. The expert's work as a technologist for the university's AI and Progress program focuses on AI regulatory design and measurement, critical infrastructure and cybersecurity, and the national AI talent base with the goal of ensuring that emerging AI technologies yield a net benefit. The expert edits and writes for Digital Spirits, a newsletter on AI policy. The expert's writing has appeared in The Hill and Noema Magazine and has been cited by the New York Times, Bloomberg, Foreign Policy, and Politico. Before joining the university, he was a research fellow at the Institute of Security, Policy, and Law in another university. He holds a BA in economics, an MPA, and an MS in cybersecurity.

**We are beginning to see AI-related framework and governance measures such as the Biden Executive Order on AI and the NIST Risk Management Framework. What are the strengths and weaknesses of such measures?**

I personally think the US approach is based on trial and error. We should not make the mistake of believing that we have a well-oiled machine because we are not super coordinated so far. The Executive Order is over 100 pages with almost every agency being called on. This Order is critical because it is a voluntary standard setting guidance which light regulatory pieces. Most of it is non-regulatory with a sort of "hands-off" approach. Following the Executive Order, various other documents are being released, either sector specific or ones that point to a smaller subset of issues. There is a standard setting guidance document in progress as well. CSIA is putting out an initial guidance on how critical infrastructure owners and operators should be approaching AI and AI risk management as an attempt to apply the NIST Risk Management Framework.

The NIST Risk Management Framework is a premier document that the US government has been spearheading, and the document is already on version 2. This shows the intention that the document is to evolve over time and match the state of risks. The best point of this Framework is that it is committed to being a living guidance that updates over time and even includes a timeline to update every two to five years. This document is also a non-regulatory guidance that helps you think about AI risks, socioeconomic factors when using AI, AI utilization, and how to govern AI risks internally. It walks through various factors to what makes a trustworthy system and use, and it implements four functions as a list of activities that organizations can follow for situational awareness. There are details on implementation and metric gathering processes as well to measure success. Again, the Risk Management Framework has a list of problems to consider and actions to follow—all voluntary. The problem with this document is that there is not much evidence on whether the Framework has made an impact. We don't know how many people and organizations are using the Framework, yet it is a good starting point to think about risks.

The US approach—from the Executive Order to the Risk Management Framework—wants people to follow the approach, and the government is taking policy action; however, many areas have unsupported mandates and provisioned asks. There needs to be a way to sort missing information, understand gaps, and have the public sector spearhead the actions related to filling the gaps.

## What else is NIST doing in the AI frontier?

NIST has its hands full laying the groundwork to AI governance. NIST is also trying to compile some use cases to help use and understand NIST Risk Management Framework applications, but there is no mandate to measure the impact and applicability of the Framework. Between the Japanese framework and NIST Risk Management Framework which may be useful to look at. About three months ago, NIST also set up [US AISI](#) with a strategic plan for red teaming and benchmarking. The US AISI is in the US Department of Commerce and was established from the 2020 AI Initiative Act and funding. Gina Raimondo, the Secretary of Commerce, is involved. The intent is for US AISI to work with the reporting requirements of the AI Executive Order on things like red teaming. A coalition of private sectors—civil society and companies—will work with AISI to inform how to do red teaming, and all the major AI labs and think tanks are involved to answer questions regarding technical aspects, social controls, and regulatory difficulties. However, NIST and the US government have never done this before, so everyone is making the approach up as we go. For example, we are trying our best to red team for certain risks, but we don't know much about the risks. Red teaming can help secure models, but we will unlikely see a restriction on AI models as stringent as those in China.

Other countries' AISIs need to think about what information the private sector needs to solve problems. The public sector should provide resources and information to support the private sector on this front. Furthermore, technical assistance models and the development process can be a good place for public-private partnership.

**[Email addition] NIST revealed "[Test, Evaluation & Red Teaming](#)" in regards to the AI EO and DoC publicized [a new NIST draft guidance from the US AISI for AI safety](#). What are your high-level thoughts and comments on these two announcements?**

At a high-level, I have two comments: there are unreasonable safety management expectations, and AISI's need to avoid cybersecurity mission creep.

Like what I expressed in our meeting, I believe a lot of the asks are beyond the scope of reasonable expectations for most private organizations. A couple points stand out. On the risk management side, I think it's highly unlikely that an organization will have the time, resources, or knowledge to adequately assess threat actors of concern. This is especially true given the broad range of misuse risks (the development of chemical, biological, radiological, or nuclear weapons, automation of offensive cyber operations, or generation of CSAM or NCII) mentioned in the report. Regarding those specific risks, it's unlikely organizations will have subject matter expertise in each risk. It's also unlikely they will have access to the intelligence needed to properly understand and mitigate each risk given classification constraints.

If managing threat actors and these risks is the concern, safety institutes will need to provide resources to make that happen. For threat actors, the institute needs to compile profiles and, to make that happen, consider a centralized information sharing and analysis center. For specific risks, safety institutes need to either provide classified red teaming or build processes that could enable the sharing of relevant risk intelligence to companies so they can properly test.

Also, safety institutes need to avoid cybersecurity mission creep. I was struck by the deep focus they place on preventing model exfiltration from organizations. While certainly related, it's my view that AI safety organizations should focus on challenges unique to AI. Cyber theft is a general challenge and therefore a tempting focus for research. Institutes need to avoid cyber mission-creep as 'yet another set of cyber best practices' is going to be counter-productive. In their planning documents, organizations should clearly delineate the bounds of their mission and what 'pieces' of cybersecurity (and other tangential challenges) are within their mandate. This will help keep work focused and effective.

## Government Agency 1

Government Agency 1 is the Deputy Superintendent, Innovation Policy at a State Department of Financial Services (DFS). The expert is a cybersecurity and technology expert whose experience spans across the private sector and government. The expert most recently served as the Senior Vice President for Cybersecurity Coordination and Advocacy at a financial company. In this role, the expert was responsible for coordinating cybersecurity matters across the company's business units and departments, as well as the company's global safety, security and technology advocacy efforts. Previously, the expert was the Senior Vice President for Public Policy and Data Protection at the company where he was responsible for policy efforts in the areas of cybersecurity and global data management, while also leading the company's industry partnerships on technology policy issues. Prior to joining the company, the expert was Director of Cybersecurity Policy on the National Security Council at the White House, where he focused on efforts to advance the Administration's cybersecurity, technology, and trade policy priorities. He also served as Chief of Staff to the U.S. Intellectual Property Enforcement Coordinator, where he helped coordinate the U.S. Government's intellectual property policy and enforcement strategies.

Before joining the White House, the expert was Counselor and Senior Advisor to the Commissioner of U.S. Customs and Border Protection at the Department of Homeland Security. In this role, he worked extensively on international trade policy and operations and global supply chain security. Earlier in his career, the expert worked in the litigation department of a law firm. The expert holds a Doctor of Jurisprudence degree and a Bachelor of Arts in political science. The expert currently serves on the executive board of the Cyber Peace Institute and the Center for Cybersecurity Policy and Law.

The interview was also accompanied by two policy managers and a policy specialist of DFS:

- Innovation Policy Manager, DFS
- Innovation Policy Manager, DFS
- Innovation Policy specialist, DFS

### **Can you describe your current work at DFS and how it is related to AI governance?**

Other than US Treasury, in the Fed, DFS, the largest supervisor in the country, particularly regarding financial services and policy. We have roughly 3,000 employees, and we oversee everyone from Goldman Sachs to much smaller financial institutions in New York, and we also focus on virtual currency and crypto issues. We also have a unique role as a dual regulator of the insurance industry,



which we work and focus on quite a lot right now. Currently, we are in the process of finalizing our public guidance on the use of AI in insurance, particularly regarding underwriting and pricing. This guidance is designed to help the insurance industry better understand the parameters we have around four areas with a focus on data inputs and outputs.

We've approached this from two angles: data inputs and outputs. First, on data inputs, the insurance industry is increasingly utilizing external data sources, going beyond traditional methods in areas like health and property and casualty insurance—insurance types that we all encounter at various points in our lives. Our primary concern is the potential for bias in the data itself. While this isn't about racism or inappropriate content, it highlights inaccuracies in the information used in the insurance process, which can be perpetuated further during underwriting.

For data outputs, we are concerned about bias in a more direct sense—specifically, the risk of creating barriers for individuals who are already underserved in this country. We fear that AI could exacerbate these issues, making it even more challenging for these individuals to access the information they need and engage with insurers. This is a significant protection concern and poses a broader challenge to the growth of the market. We believe our guidance will be the first of its kind to address these categories comprehensively.

Additionally, we are developing guidance on AI-related cyber threats and exploring how AI can be used to mitigate these risks. We want to provide a balanced view, showing that while AI poses challenges, it can also offer valuable solutions. We recognize that a variety of enhanced cyber issues are emerging, particularly related to social engineering and misinformation. These threats impact various areas, including election integrity and the oversight of financial institutions, which are increasingly targeted and vulnerable. Additionally, there are concerns about supply chain risks and third-party dependencies as companies integrate more AI into their operations. What's particularly interesting is that we believe technology is reaching a tipping point. In the next two to five years, we anticipate significant advancements in real-time AI generation that will create highly engaging content, including video. This evolution is on the horizon and will change how we approach these issues.

**How does the DFS coordinate with other bodies in the United States (whether at the state or federal level)?**

At the state level, we are working with counterparts in California, the Department of Financial Protection and Innovation (DFPI) on a variety of issues. At the national level, we are engaged with the CFPB, the Federal Reserve and the Treasury.

**The Treasury recently released a report on [‘Managing Artificial Intelligence – Specific Cybersecurity Risks in the Financial Sector’](#), what is your perspective on how this impacts the work you’re currently doing?**

I think they did an excellent job with that report; it really sets a marker for many state regulators. While the Biden Administration's EO 14110 covers a broader range of topics, the Treasury's actions are among the most important and specific. My understanding is the EO called out for the departments and agencies to do those sorts of stuff and the Treasury was one of the first to respond.

**In the context of AI governance, how important do you think risk management frameworks are compared to more prescriptive regulations?**

In our work, we prefer to provide guidelines. Notice on our work around the insurance industry and cybersecurity issues, we talk about guidance rather than regulation and we took a purposeful approach here. Instead of being overly prescriptive at this stage, we can leverage our existing laws and regulations to provide clear but flexible guidance to our regulated institutions. We have successfully implemented this approach in both cases, and I fully support it. At some point, we may need to establish regulations for valid reasons, but particularly in the context of AI, I find this approach crucial because the field is evolving so rapidly. As regulators, we must avoid creating barriers that could stifle innovation. Also, on the macro approach, we take a very proactive stakeholder engagement approach in terms of regulations and guidance. For instance, our teams spent over a year focused on the insurance guidance, engaging both informally with industry experts and formally through public comments after we released the initial guidance. It's rare to see proposed guidance include public comment periods, as this is typically more common in regulatory contexts.

**Which agency seems to have central role (i.e. role to implement regulations/guidelines) from enterprise viewpoint? CISA? MITRE?**

I believe CISA prefers frameworks and guidelines rather than prescriptive regulations. However, I have limited insight into the institutional relationship between CISA and MITRE in developing these guidelines. Relatedly, on non-government efforts, I know the OWASP Top 10 has played a role in developing a set of issues around AI and cyber.

**Japan, so far, does not have a specific AI office or department yet and all AI-related regulations have been broad and overarching. What do you think about this approach? Do you think there needs to be a specific AI office or department?**

I recently had a discussion with the Superintendent after our first meeting, where we established an AI steering committee that includes our executive leadership teams. This committee, which I chair, will meet every two months to discuss emerging issues across our divisions. We will identify areas that may require departmental actions or regulations. I don't believe it's necessary to have a Chief AI Officer for every division. Instead, we should focus on an enterprise-level approach. Everyone is grappling with the challenges and opportunities that AI presents. I've observed this in both corporations and government; a steering committee approach is more effective than appointing a Chief AI Officer.

In organizations, it's important to involve various roles such as the Chief Data Officer, Chief Privacy Officer, CTO, and CIO. I anticipate that as we move up through government levels, we will see similar debates as we did in cybersecurity. While we have Chief Information Security Officers (CISOs), there isn't a dedicated Department of Cybersecurity. Instead, we rely on the improving interagency processes. As part of our stakeholder engagement approach, our steering committee will invite an external expert on AI for each meeting. The next session will feature a guest speaker who will brief us for about 45 minutes to an hour, followed by internal discussions. This approach is crucial for building

expertise and comfort within our agencies as we develop policy work. Engaging early with knowledgeable individuals is critical, especially on this issue. Early engagement also helps shape the government's role and develop effective policies, while also fostering buy-in for those policies.

**Recognizing that the steering committee just met for the first time and will soon hold a second meeting, what are your thoughts on which expert to bring in or which topics to focus on. Do you envision sticking to more specific issues, or something broader?**

Actually, while specific topics are important, I believe we should emphasize the use cases of AI within the service sector. This is crucial because AI is already impacting regulated industries and could create barriers for small and medium-sized businesses. We saw similar challenges in the cybersecurity space, where under-resourced and understaffed small and medium businesses struggled to compete.

The divide is even greater now compared to previous years; it feels like a significant gap between 2024 and 2014 in terms of cybersecurity access. Today, small and medium businesses face greater difficulties in accessing effective cyber defenses, despite increased awareness since 2014. The tools and cost structures for cybersecurity have become much more complex and expensive. I believe this challenge is already emerging in the context of AI, and as a regulator, I'm quite concerned about it. Small and medium-sized businesses are vital contributors to economies at every level—city, county, state, and country—both from a revenue perspective and in terms of fostering innovation and competition.

**What is a factor that should be prioritized when developing AI governance measures?**

I would say governance itself is key—having board and senior management oversight related to risk management and audit is crucial. It shouldn't be overly complicated to apply those governance principles to the AI context. We've seen significant collaboration in recent years, particularly with the SEC on cyber requirements related to board expertise. There was a notable gap in understanding cyber issues at the board level, and I believe we'll encounter similar challenges with AI governance. So, for me, governance is fundamental.

Policy manager: I think it's essential to communicate effectively with those new to this field and have a board that knows more about it. Bringing in expertise from other areas of policymaking can be beneficial.

Policy manager: I believe that effective policy helps drive market competition and pushes larger market participants to take on the cost of complying in order to avoid the risks.

Policy manager: Regulatory risk is a significant concern for large corporations. Regulating AI is not so different in what you do for model management and internal processes. The industry has received many warnings, such as from the Bank of England, emphasizing that if you can't explain your model, it might be best not to use it at all. It's crucial to understand what regulators expect from the industry to deploy models effectively. Establishing a feedback loop with regulators in-house can help facilitate this understanding.

**Both Japan and the US are trying to figure out testing mechanisms for AI safety. Some are engineering based, some are social sciences based, but this is very difficult depending on viewpoints. What is your insight into AI testing, how important is it?**

I think it's hard to do. There are some cases where there are life safety impact elements where you could probably have stronger guidelines that are clearer. However, addressing socio-economic developments is much more challenging. Take TikTok as an example. The platform's algorithms are perpetuating certain perceptions, particularly around issues like age and identity. This week, we saw a specific instance where its AI is shaping narratives that could influence public opinion and potentially impact elections. The question is: how do we implement AI-specific guardrails in cases like this?

**On your point about AI perpetuating narratives, or in some cases fake news, in democracies it is hard to prescribe what crosses the boundary and therefore address it. What is your perspective on how to approach this?**

The challenge we face with fake news mirrors our issues with fairness and defining truth. If you believe in free speech, it's tough to determine what crosses the line. However, there are safe ways to address these concerns. For example, our team tried using an insurance model with a "five-foot data" approach, examining both data inputs and outputs. Each had different testing components, focusing on outputs in the context of U.S. law, particularly regarding disparate impact. We asked whether there are less discriminatory alternatives that are still effective. I believe we can borrow from other technological regulations and standards rather than relying on a single set of AI guidelines. Integrating different elements is essential; otherwise, it becomes challenging to navigate.

Policy manager: This complexity resembles climate change discussions, and ESG frameworks, as firms now view these issues as enterprise wide. They want to implement solutions across all areas.

### [Election Analyst 1](#)

Election Analyst 1 serves as counsel for the Brennan Center's Elections & Government Program, where the expert's work focuses on election reform, election security, governance, voting, truth and information. The expert was previously an attorney with Earthjustice, a nonprofit organization, where she engaged in a wide range of federal litigation and policy work. The expert served as counsel for Native American Tribes challenging the U.S. Army Corps of Engineers' approval of the Line 3 pipeline and was counsel for the NAACP in a lawsuit seeking stronger federal standards for lead in water. The expert also served as lead counsel on a federal Civil Rights Act matter, successfully guaranteeing the right for Spanish-language and other immigrant communities to participate in environmental decision-making in the state of Texas.

Prior to Earthjustice, the expert was an Equal Justice Works Fellow at the Natural Resources Defense Council where the expert helped lead the organization's work on post-disaster environmental and governance issues in Puerto Rico and contributed to efforts to abate severe drinking water pollution nationwide. The expert's writing has been published in the New York Times, the Washington Post, CNN Opinion and NBC News. The expert has co-authored nationally recognized reports and the

expert's writing has been entered into the Congressional Record in Congress. The expert holds a J.D (Juris Doctor).

### **How have election-related threats evolved from 2016 Cambridge Analytica Era to the 2020 Twitter Era to the 2024 Generative AI Era?**

Generative AI exasperates pre-existing risks. Disinformation existed prior to generative AI, and the growth and explosion of social media have changed the landscape. Microtargeting certain populations to spread information faster and more widely is happening more. Generative AI changes the speed, scale, and scope of this kind of disinformation. It's also cheaper to produce more sophisticated content now. Threat actors can combine generative AI with other kinds of AI like bots to perform very sophisticated microtargeting which can be particularly problematic.

### **How would you describe the severity of AI-enabled disinformation and information operations in the 2024 elections around the world?**

We must see how it plays out, but in the US context, we see less deployment of generative AI by campaigns and that is in part because there is an emerging norm driven by laws and policy discussions on more careful and thoughtful use of generative AI. Outside of campaigns and political committees, generative AI is being created and spread by private actors, but the impact is not clear yet. An underexamined risk is the use of large language models (LLM) in disinformation campaigns. We have seen encrypted platforms like WhatsApp spread disinformation that impact Indian American communities and Latino communities. Threat actors use LLMs to produce continuous interactive conversations with voters through AI-generated robocall or messages. AI is responding to voters and adapting to the situations, and sometimes we don't have enough visibility into these platforms to find evidence. Historically, we have seen cyberattacks on election infrastructure as well as phishing attacks against election officials. Generative AI can exasperate these risks as well as generating phishing emails will get easier and faster. Voice synthesis AI also increases the risk. Phishing attempts have increased over the past couple of years.

### **What are the most effective methods for mitigating AI-enabled disinformation and information operations?**

We have had several states pass deepfake laws in the US. There are also bills in consideration, but passing laws on technology in the US is difficult. The laws that have passed generally fall into two categories: laws that focus on visual or audio generative AI-created deepfakes. There is little focus on LLMs, but political communications from campaigns and committees underscore deepfakes a lot. The Brennan Center did tabletop exercises with election officials at the state and local level to help them run through AI-related threats and risks to game out potential responses. This sort of exercise is also helpful. We published [recommendations](#) online as well.

### **How is marking AI-generated content being approached?**

The Biden Executive Order instructed federal agencies to create standards around content provenance and that will be used in the context of federal agencies using AI, but the idea was to create a blueprint

for states and local governments as well for providence standards and information verification for official content. Watermarking and providence standards are things we find critical but complicated to implement. The California bill has done work with this element: establishing watermarking standards as well as labeling certain chatbot conversations.

**Is most AI-enhanced disinformation domestic or foreign in origin for the US?**

Both in the US. Microsoft has reported on China being an actor in this front, and Meta reported that the disinformation in the European Union election was mostly domestic in nature.

**Will generative AI make us question what is fundamentally true versus fake?**

There is the liar's dividend for sure as seen in the War in Gaza. This is the idea that true or authentic content becomes easier to smear as fake, and this will affect elections across the world as well. People will definitely distrust authoritative information more, and information will be easier to discount. Thus, there is imperative to introduce digital literacy initiatives.

**What are some use cases of AI in election that we have seen thus far or will see—both the good and the bad?**

Non-generative AI has been used in elections before. Election administration uses AI to maintain voter registration databases, verify mail ballot signatures, and more. These uses support election administration, but they also create produce risks, meaning there needs to be regulation. Officials use generative AI to simplify language and edit documents to make them more accessible. Officials also are interested in language translation, but these uses need to have guardrails in order to mitigate potential inaccuracies. We need to use AI as a supplement, not a substitute. Generative AI can also help under-resourced campaigns compete more effectively and help level the playing field.

There is potential for interactive disinformation campaigns driven by LLM is troubling and need to be addressed through regulatory mechanisms. Older populations and certain demographic groups are vulnerable to this technique. Finally, some cybersecurity risks are exasperated with sophisticated spear phishing attempts that use generative AI-produced visuals and audio.

**What are some laws and tools that can help regulate AI risks?**

There is now a stronger push to have a comprehensive privacy law to mitigate AI risks. Regulation focused on platform more than the user could also be helpful. Generative AI detection tools work sometimes, but there are flaws and inconsistencies. Using fact checkers that are transparent and have information on their confidence level of analysis are helpful as well.

## AI Policy Expert 5

AI Policy Expert 5 is a computer scientist who most recently served as the first Director of Artificial Intelligence for New York City and is currently Adjunct Associate Professor in the School of International & Public Affairs at a university, where he teaches a course on AI for policymakers. Previously, he co-founded a technology startup, which was acquired after 10 years in operation, was Inaugural Fellow at a policy incubator of a think tank and worked as a senior quantitative analyst at an investment bank. He received his Ph.D. in computer science, focusing on large-scale machine learning and convex optimization. His research has received over 20,000 citations in academic literature and is widely used in industry.

### **What was the approach for and focus of the [NYC comprehensive Artificial Intelligence Strategy](#)?**

Ultimately the NYC Comprehensive Artificial Intelligence (AI) Strategy provides findings and opportunities across different themes. The first purpose was pedagogical. There is pedagogical content about AI itself, which is not specific to New York. That's separated out in an appendix that we released separately, called the New York City AI Primer. This Primer can be used beyond NYC any government, business school, corporate leadership, or activist organization. It is around 30 pages long and is intended to be readable by practitioners as well as senior executives and senior agency officials. It has two parts: one explaining how AI works and another discussing ethics, governance, and policy. The document goes into more detail about what we mean when we say something is unfair and how that can manifest. That's the core AI part. Then there's the ecosystem component, where we mapped out five thematic areas. These areas are partly technical and partly governance:

- Data Infrastructure: substrate on which all machine learning and AI is built.
- AI Applications: cases where an agency might be using an AI system.
- AI Governance/Policy: applications within public and private sectors that the city regulates.
- Partnerships: discussion on how NYC could/should partner with stakeholders for projects.
- Business and Economic Development: labor workforce issues such as talent pipelines.

In the ecosystem mapping component, we identified all the different agencies in the city government that relate to any component in these five thematic areas, resulting in dozens of different agencies. Some agencies use AI but are not part of its governance, and vice versa. We also conducted an exercise to map out the different civic institutions and city institutions that relate to AI. These institutions are diverse and include everything from community boards and neighborhood associations to industrial research labs. An industrial research lab might be housed in a large tech company, but it is distinct from the company itself, with different priorities and roles in the city infrastructure.

We completed this ecosystem mapping and included a section on findings, where we examined what has been happening in each of these thematic areas related to the overall strategy. Many people did not necessarily think of their work as AI-related, but it can be viewed in that light, and we identified various opportunities for further development in each area. All of this was informed by a substantial number of interviews conducted with individuals from both the city government and the broader community, including corporations, non-profit leaders, think tanks, and universities. This is documented in another appendix of the report.

This report was released near the end of the de Blasio administration. It is deliberately non-prescriptive about next steps; instead, it aims to identify various opportunities and provide a conceptual framework for the subsequent administration to build upon. The goal is to allow them to establish their priorities while working within the existing framework. The new administration has pursued this approach and the AI Action Plan, which outlines more concrete strategic and tactical objectives. This document is not a pedagogical framework but rather a more action-oriented initiative.

### **What are the criteria for effective AI governance measures?**

The AI Strategy avoids being prescriptive about certain aspects. Some of it is simply factual, highlighting what governance measures should consider such as the diversity of applications, agencies, purposes, and so on. Even though the strategy does not specifically mention large language models (LLMs), nothing in there is made irrelevant by them. While there may be additional aspects to consider, LLMs or generative AI do not contradict anything in the strategy. Some people have attempted to create very prescriptive frameworks, but a significant mistake people make is trying to adopt a one-size-fits-all, monolithic approach. This will never work. Whether it's the NIST AI RMF or any other similar system, it simply doesn't work.

In the second findings and opportunities section, we go through six AI applications in city government and make it clear that there are qualitatively different kinds of applications. Even if you're not involved in AI, it should be obvious that any governance measure applicable to one application would not necessarily apply to another. For instance, one application involves criminal justice, which directly affects people; the input is a person in a case. They conducted significant community engagement and interviews with people in the criminal court system regarding the criminal justice application. In contrast, this approach would be completely inappropriate for a cyber defense system that operates behind the scenes on city infrastructure. This system does not interact with people at all; it functions automatically hundreds of thousands of times a second. Issues like social fairness are irrelevant here.

Thus, having a single governance framework would result in something so vague that it wouldn't be useful as it would need enough flexibility for those enforcing it to selectively apply or ignore parts based on the specific application. The questions that arise are very application- and domain-specific. I think it's critical to start from the bottom up—pick a clearer area that has relevance and work through that before expanding further, rather than adopting a top-down approach with a broad AI governance framework that will ultimately fail.

Another important consideration is the purpose of the governance measure. Many governance measures lack a clear stated objective. For example, the AI registries that some governments have developed require substantial effort to catalogue. Advocates pushed heavily for this, but interest has waned after governments produced these registries. While I find them useful for teaching—providing examples of societal applications—they are not worth the effort. Some of this initiative parallels the open data movement, but open data is different because it has tangible uses; various agencies, researchers, and companies utilize it. In contrast, these registries often serve merely as disclosure forms. For instance, the New York City Department of Health has a machine learning model that analyzes Yelp reviews to serve as an early warning signal for food safety. This system is straightforward



and flags restaurants where people report feeling sick. While it's interesting to know about, it doesn't raise significant government secrecy, privacy, or disclosure issues.

So, why establish a huge infrastructure to disclose such trivial information? Much of this was motivated by a paranoid view that advocates believed AI was being used in numerous controversial contexts. They anticipated that disclosing these systems would reveal many contentious applications. However, the reality is there are only about 20 applications, most of which are mundane, and the more interesting ones are generally known to those in the relevant departments. I would suggest starting with a clear understanding of the purpose of any governance measure rather than jumping on the bandwagon of doing something about AI simply because others are doing it.

**How effective do you think AI safety practices measures by NIST and CISA are in suppressing misuse of AI?**

I don't think they do much. One thing is that it's not mandatory to use. The people who choose to use it will likely be conscientious individuals with applications that are irrelevant. For instance, there might be someone from the parks department saying, "I did the RMF," but they're just addressing how busy each park is so that people can go to the less crowded playground. It's great, and it's a perfectly nice application. However, the RMF is not really designed for this. More concerning departments likely have their own systems, and it's important to avoid balkanization. There are initiatives at NIST, the White House, the National Telecommunications and Information Administration (NTIA), and the Department of Defense (DOD), among others. Are these entities all communicating? Not really. When push comes to shove, is any project likely to be delayed for compliance? Likely not. Are they going to refuse to procure a system because of it? Probably not. In cases where they might, those are not typically high-stakes situations. The controversial systems, like automated weapons systems, will proceed as planned. There needs to be real enforcement in regulation to prevent issues like that. There are significant competitive pressures at play, both interagency and from other priorities.

I don't think it's an easy task, but I would try to build on something like cybersecurity, where people have internalized the understanding that, even if it causes delays in rolling out some projects, it's just too risky to implement insecure systems. Part of the problem is that some AI safety and ethics concerns have reached a similar level of urgency, while others have not. By pushing the narrative about AI ethics, largely driven by activist groups, we've seen some useful outcomes. However, it has also started to backfire, undermining the importance of genuinely safety-critical systems. One critical aspect is simply whether the system functions as intended. Sometimes you buy an AI system, and it doesn't perform as claimed, or it works initially but degrades over time due to model issues or other factors. I would suggest building on frameworks that already work. However, whenever you introduce more red tape, more bureaucracy, more delays, or increased costs, you need to consider how these conflicts with other priorities will be resolved and who will make those decisions.

**Despite the lack of mandatory regulation in cybersecurity, how did organizations get to the point of internalizing the need to comply voluntarily? How does this apply to AI?**

Constant huge disasters. There are incidents like the AT&T breach and other significant breaches that introduce substantial financial and reputational costs for the company. When you have one of these

breaches, a whole series of internal actions must take place. You might even get called in front of Congress. Even then, people cut corners. They try to avoid it because it creates more overhead. The issues with AI are much worse because, while cybersecurity is hard, it is much more rigorous. They have proofs in place—essentially, cryptographic protocols that ensure security. AI, on the other hand, is not like that. There are a few areas where you might have a more rigorous understanding of correctness and ways to test for it, but when it intersects with socio-political issues, it becomes subjective and debatable. There is no definitive answer, and you must make a value judgment.

AI governance is about trade-offs. Officials must accept that there is no way to satisfy all objectives. If you increase the fairness of a system, you may render it unusable in situations like medical diagnosis. You don't want to compromise on medical diagnoses' accuracy either though. Similarly, there are trade-offs between fairness and privacy. One reason for this unfairness could be the lack of data on minority groups, but collecting more data can be perceived as surveillance. I've seen this happen in the Bronx. These issues are fundamentally in tension with one another, and the governance framework must acknowledge this at a deep level. This is why the domain-specific approach is so critical. The way these trade-offs are navigated will necessarily differ in each situation. In some cases, privacy is a huge issue; in others, it's not an issue at all.

### Data Consultant 1

Data Consultant 1 is a policy and tech translator, product consultant, and long-term digital strategist guiding the intersection of emerging technologies, culture, governments, and policy. Equipped with degrees in both computer and cognitive science, the expert focuses on data governance, data security, artificial intelligence (AI), and privacy in the digital age. The expert is a subject matter authority who has written extensively about AI and other data driven topics for over a decade. The expert is also a member of the *Washington Post's* The Network, "a group of high-level digital security experts" selected to weigh in on pressing cybersecurity issues.

An information privacy professional, the expert has led standards and policy efforts around emerging technologies throughout the expert's career. The expert regularly collaborates with stakeholders and policymakers in Washington, DC, and with global product and policy teams, to help steer the conversation on the role of AI in society and its impact on privacy and security. The expert previously served as the director of the privacy policy team at one of the world's largest social technology companies, leading policy development around the company's work to develop privacy-protective product experiences, building policy frameworks that create accountability, and promoting privacy-protective decision making across the company.

Earlier, the expert helped found the public policy team at a website performance and security company and served as the global and federal privacy and security issue expert on a multinational technology company's public policy team. The expert started the career working on government technology, privacy, and identity management at a public interest group focused on the rights of individual users in relation to technical policy.

**We are beginning to see AI-related framework and governance measures such as the Biden EO and the NIST AI Risk Management Framework. What are the strengths and weaknesses of these AI regulations and frameworks?**

The AI risk management framework is robust for several reasons, notably its flexibility and applicability across various use cases. It recognizes that risk is highly context dependent. By addressing questions not only about development but also about how AI systems are used, it enables much better governance compared to some other frameworks we've seen proposed. One of its key strengths is that it builds on existing knowledge around privacy and security from established frameworks and global discussions, rather than treating AI as entirely novel. Many have approached AI as something completely new, attempting to reinvent the wheel, but that perspective is fading as similarities become more evident. The AI RMF was ahead of its time in recognizing that much of what we need to do to manage AI risk closely resembles practices for other applications.

However, one limitation is that it can be quite technical and detailed, which may be challenging for those not deeply familiar with the subject. While it serves as a useful guide for practitioners seeking detailed insights, it may be less accessible for policymakers who prefer higher-level information that is easier to understand and promote as a viable solution. I'm a big fan of the AI RMF. Its multi-stakeholder approach has been notably successful, fostering goodwill among industry players who are actually building these tools.

Moving on to the AI EO, I find it impressive. The White House's work has produced a range of individual, discrete actions, some more detailed than others, including studies, reports, and concrete initiatives. Each action can be assessed individually, though I won't attempt to cover them all here.

The sheer breadth of the executive order is both a strength and a potential weakness. While all actions are productive and contribute to progress in various ways, their impacts will vary significantly. For instance, Section Four, which focuses on security and safety, has yielded concrete outcomes, particularly concerning advanced AI models—referred to as frontier models and dual-use foundation models. This section has demonstrated swift results from a governance perspective. In contrast, the civil rights components address existing laws or highlight the absence of comprehensive privacy legislation, relying instead on individual sectoral laws. This unevenness exists, but overall, I find the ambitious set of actions commendable, and I'm impressed with how the administration has consistently moved these initiatives forward.

**Are there specific priorities within the EO's 100+ actions that you consider most important? Could you identify the key sections or initiatives that stand out to you?**

To clarify, my personal focus is on making AI a secure, safe, and trusted tool that can be widely used across various entities. This perspective informs my priorities. Within Section 4, I see several high-priority elements, though they may not align with the Department of Commerce's Bureau of Industry and Security (BIS) pieces that are concentrating on regarding dual-use foundation models. I believe that the work NIST is doing to clarify and expand on various frameworks, including the AI RMF and the Secure Software Development Framework (SSDF), is incredibly important for advancing this initiative. While I recognize that these are voluntary frameworks, they provide essential guidance for individuals in day-to-day roles who are evaluating whether to adopt this technology. These tools will significantly enhance people's comfort and confidence in using AI.

**Do you see the development of AI risk management evolving in a way similar to the cybersecurity framework, particularly in terms of becoming a global standard through voluntary adoption and private sector input?**

I completely agree. I believe that cybersecurity was somewhat easier—though that’s a complex statement—to internationalize. In contrast, AI has become much more polarized and politicized. That said, I truly hope that the AI Risk Management Framework (RMS) can play a similar role to the cybersecurity framework. Having wildly divergent frameworks around the world doesn’t benefit anyone; it risks fragmenting the technology in ways that are unhelpful. I think there’s arguments that there are differences in what every region or country needs, or even community. There are a lot of ways to break this down, but I do think that coming up with a largely consistent framework would be incredibly helpful to everyone.

**What do you believe are the factors or criteria needed for effective AI framework/governance?**

I believe several key aspects are necessary for these frameworks to be effective and impactful. Firstly, they must be risk-based, which surprisingly seems to be a controversial point. For instance, the EU AI Act aims to be risk-based, yet it includes provisions that are not risk based. Some elements declare certain uses off-limits regardless of the context or actual risk involved, while other sections of the act are indeed risk-based.

Effective frameworks and governance must be contextual and risk-based because these tools can be used for a wide range of purposes. You cannot impose a one-size-fits-all governance model on a general-purpose AI system. For example, using a language model to summarize song lyrics is fundamentally different from summarizing medical records for healthcare professionals—both tasks may seem similar but carry vastly different risks. Thus, the emphasis on being risk-based and contextual is crucial. Any framework or governance model must be grounded in practical realities; any governance model needs to function effectively in real-world scenarios rather than remaining at a hypothetical level. High-level, non-contextual governance simply doesn’t work.

**How do you view the challenge of placing the burden on enterprises to contextualize their use of AI and assess associated risks? What steps do you think are necessary to ensure that these organizations have the right expertise to conduct effective risk analyses?**

It’s challenging, but there’s room for figuring out what can be specified. When I refer to the supply chain, I mean the entire deployment process—from development to implementation. We should assess whether adequate governance was in place throughout the process, while recognizing that some evaluations need to occur at the end. If I step back and consider my biggest concern regarding AI, it’s not that super powerful AI will malfunction and pose a threat to humanity. Instead, I’m more worried about the misuse of AI that isn’t suited for critical tasks. For example, if you were to ask ChatGPT to manage a waste management plant, it would genuinely try its best, but it simply isn’t designed for that role. It’s crucial to ensure appropriate use of technology. While there are general-purpose models that could be suitable, ChatGPT isn’t one of them. I don’t know how to move that up, frankly, the expertise chain, where people are more knowledgeable about the technology and the

development of the model itself. I do think that's something we can figure out and I do point at explainability a lot as a tool that can really help here.

**Can we draw lessons from other sectors, like financial services, to gain better control over the models they use, and how can we effectively integrate AI into model management systems?**

I agree, there are a few sectors that are ahead of the curve. Financial services, for example, have been working with complex models and have been regulated. In contrast, the agricultural sector, for instance, lacks that level of expertise.

**What are the challenges in implementing the NIST AI RMF compared to the NIST Cybersecurity Framework?**

There are very few areas where the AI RMF and the cybersecurity framework significantly diverge in their implementation and learnings. The key point is that the cybersecurity framework benefits from a much broader knowledge base. During a panel I moderated with several experts using the AI RMF—some of whom were consultants from firms like EY—everyone agreed that the biggest challenge was that participants had jumped in with enthusiasm but were starting from a point much further ahead than their audience. They realized the need to rewind and help build the foundational knowledge required to effectively implement the AI RMF. This involved level-setting to avoid assuming that everyone is at the same stage with both frameworks. It's crucial to assess where people currently are and then identify how to bridge that gap. We're now seeing this massive knowledge transfer from folks in IT towards folks in business and leadership and that is going to be process.

**What are US departments and agencies' roles in regulating AI-related risks?**

I think one of the reasons AI is already being effectively regulated in the US, particularly in certain sectors, is due to our existing technology neutral laws. These laws allow various agencies to implement regulations without Congress explicitly granting them authority over AI. Obviously, the financial sector is one major example. think the health care sector, eligibility, and civil rights, all of those are tech neutral. I believe that this approach is significant, and I support it. However, an AI regulator will face substantial challenges in coordinating across agencies. As you mentioned, these agencies often don't collaborate effectively. For example, the FTC and CFPB are both very eager to regulate AI, and they frequently compete over their respective roles, particularly in the consumer-facing financial sector, which both view as their territory.

**One of the challenges of the US approach to cybersecurity and AI regulation is the existence of multiple independent regulators and quasi-coordinating bodies. Based on the US model, what challenges might arise for the Japanese AI Safety Institute?**

I do think it will be a challenge for the Japanese AI Safety Institute (AISI) is to define the boundaries of regulation. I can't think of anything I use daily—aside from maybe a water bottle—that doesn't have some AI aspect to it. For instance, my chair doesn't, but many everyday items, like a smart heating mug for coffee, clearly incorporate AI. This integration makes it difficult for an agency to regulate AI effectively across various use cases. If that's the structure in place, the biggest challenge will be ensuring that the AISI collaborates well with others and avoids the pitfalls seen in the US

administrative structure. I would love to see improvements in that area, but figuring out how to foster effective collaboration will be a significant hurdle.

**What is your perspective regarding the responsibilities of developers versus users in AI governance?**

I think there are three categories to consider: AI developers, deployers and integrators who build AI into tools, and end users. If you focus on only one category and overlook the other two, you won't effectively manage the overall approach.

**You've written about the role of a Chief AI Officer within federal agencies, where do you think this role should report within an organization for optimal oversight and collaboration?**

This project is ongoing, and one reason we wrote that blog post was to gather input from those with strong opinions. Personally, I believe we need a slightly different structure than what currently exists, particularly with the need for a chief risk officer. Currently, the chief AI officer often falls under one of several departments: the CISO's office, the CIO's office, the chief product officer's office, or the general counsel's office. Each of these placements has its shortcomings. On the other hand, many organizations are forming kind of AI councils to address coordination issues across these groups, but I feel there should be a designated individual responsible for this area. The chief AI officer's role, especially in terms of addressing risk, should be integrated into a broader risk management structure rather than just an AI management structure. While it remains to be seen if we can develop new and interesting ideas, I'm increasingly convinced that all governance and risk management roles should roll up to one person overseeing risk—not solely from a legal perspective.

## 10. Appendix

### Appendix A: Case Studies of AI in Recent Elections

Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
Taiwan	2024	<p><b>Chinese Disinformation Campaign in Taiwanese Elections:</b><sup>283</sup></p> <ul style="list-style-type: none"> <li>• Falsified political documents and incidents, including DNA tests and "hacked" military documents, to influence the Taiwanese election.</li> <li>• Used AI-generated avatars to amplify disinformation campaigns, particularly against DPP candidate Lai Ching-te, accusing him of corruption and embezzling military funds.</li> </ul>	AI-generated fake avatars	<ul style="list-style-type: none"> <li>• Despite efforts to sway voters, there was no significant impact on the election outcome.</li> </ul>
India	2024	<p><b>Deepfakes in Elections:</b></p> <ul style="list-style-type: none"> <li>• A viral video featuring a Bollywood actor mocking the Bharatiya Janata Party (BJP) for not fulfilling election promises and calling for support for the opposition party emerged.</li> <li>• Later the actor denounced the video as a deepfake.</li> </ul>	AI-generated deepfakes	<ul style="list-style-type: none"> <li>• Deepfakes and voice cloning have contributed to increased costs.</li> <li>• Increased production of digital content requiring growing investment in technology to check emerging risks faced by party leaders.<sup>284</sup></li> </ul>
India	2020	<p><b>Deepfakes in Elections:</b></p> <ul style="list-style-type: none"> <li>• The BJP party released manipulated videos featuring one of their party's elected officials speaking in two different languages to appeal to different electoral blocs.</li> </ul>	AI-generated deepfakes	<ul style="list-style-type: none"> <li>• The party defended the use of deepfake videos in this case, but public criticism was heated, contributing to anxieties over the use of GenAI technology.<sup>285</sup></li> <li>• Though experts have described the potential for such use cases to help reach voters in far-flung areas and from various social and linguistic backgrounds, they are also concerned over the transition from election strategies which emphasize physical rallies and speeches toward an environment in which digital narratives are more important.</li> </ul>

Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
Indonesia	2024	<p><b>Deepfakes in the Election:</b> <sup>286</sup></p> <ul style="list-style-type: none"> <li>• The Indonesian General Elections Commission took a relatively hands-off stance about AI-generated content used in campaigns. The country had been plagued by deepfakes ranging from fabricated health claims supposedly made by the former Minister of Health to a political deepfake video of the late Indonesian President Suharto urging voters to vote for Golkar candidates. However, there has been no evidence of widespread disinformation campaigns in the election of Prabowo Subianto.</li> <li>• There is high social media penetration in Southeast Asia, making it a fertile ground for AI-driven disinformation.</li> </ul>	AI-generated deepfakes and news stories	<ul style="list-style-type: none"> <li>• Improved public literacy around AI, empowering media practitioners and journalists to share tools, skills, and implementing practices in combating disinformation are all mitigation tools that experts explain can be utilized to address concerns.</li> <li>• Integrating AI policies with the interests of the electorate is another suggested solution for future issues. <sup>287</sup></li> </ul>
US <sup>288</sup>	2024	<p><b>Disinformation in the New Hampshire Primary:</b></p> <ul style="list-style-type: none"> <li>• AI-powered robocalls impersonating US President Biden circulated discouraging Americans from voting.</li> <li>• There is a misconception that these attacks only influence voters from less-educated backgrounds. This contributes to the “dumb voter” trope, which marginalizes certain populations and contributes to the appeal of populist candidates.</li> <li>• Incidents involving CEOs of large companies falling prey to voice-generated deepfakes have also emerged in recent years, demonstrating that similar threats endanger voters of all social classes and educational backgrounds.</li> </ul>	AI-generated robocalls	<ul style="list-style-type: none"> <li>• While a smaller fraction of voters participates in primary elections, primary elections decide who appears on general election ballots.</li> <li>• GenAI used to exacerbate confusion about the electoral process and influence outcomes.</li> <li>• Safeguards targeting the use of such technologies are necessary, including equipping voters with more robust media literacy skills.</li> </ul>



Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
US <sup>289</sup>	2024	<p><b>Biden Death Rumors:</b></p> <ul style="list-style-type: none"> <li>• Rumors spread on X in July purporting that President Biden was mortally ill or already dead.</li> <li>• No information supported these claims, but X recommended posts to users that failed to question the veracity of these claims.</li> <li>• When the platform was known as Twitter, trending topics and posts were vetted by humans to curtail disinformation, but since Elon Musk’s purchase of the platform in 2022, summaries of trending topics are now provided by the platform’s AI software.</li> </ul>	AI-vetted Disinformation on X	<ul style="list-style-type: none"> <li>• Experts have called into question the neutrality of the X platform and its vetting processes for information featured in trending topics.</li> </ul>
US <sup>290</sup>	2024	<p><b>AI-Generated Fashion Show Featuring World Leaders:</b></p> <ul style="list-style-type: none"> <li>• Social media platform X’s owner Elon Musk shared an AI-generated video featuring world leaders walking in a fashion show.</li> <li>• The video included images of Joe Biden in a wheelchair and Donald Trump in a suit.</li> </ul>	AI-generated deepfake	<ul style="list-style-type: none"> <li>• The video went viral and sparked media attention around the world. False images of world leaders may contribute to biased perceptions or more broadly to “Liar’s Dividend” threats.</li> </ul>
US <sup>291</sup>	2024	<p><b>Biden Cursing Videos:</b></p> <ul style="list-style-type: none"> <li>• An AI-generated video featuring President Joe Biden cursing spread across social media platform X following his official announcement to drop out of the 2025 presidential race. The clip appears to show Biden using anti-LGBTQ slurs and curse words.</li> <li>• The video featured the logo of American public broadcaster PBS.</li> </ul>	AI-generated deepfake	<ul style="list-style-type: none"> <li>• The viral video prompted the media to disseminate alerts that the speech from Biden was not real.</li> <li>• PBS denounced the video as a deepfake and opposed the use of misleading fake videos.</li> <li>• Could broadly exacerbate the “Liar’s Dividend”.</li> </ul>
US <sup>292</sup>	2024	<p><b>Anthony Hudson Endorsement Video:</b></p> <ul style="list-style-type: none"> <li>• Michigan Republican congressional candidate Anthony Hudson’s TikTok account posted a deepfake video featuring the voice of Martin Luther King, Jr.</li> <li>• The impersonation included an endorsement of the Republican congressional candidate.</li> </ul>	AI-generated deepfake	<ul style="list-style-type: none"> <li>• Candidate Anthony Hudson responded that a volunteer posted the video without his campaign’s knowledge.</li> <li>• Several democratic politicians expressed dismay at the video, calling into question the Republican Party and its supporters.</li> </ul>

Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
US <sup>293</sup>	2024	<p><b>AI Chatbots and Election News:</b></p> <ul style="list-style-type: none"> <li>Recent reports indicate that when asked about basic election information—such as polling locations or voter registration requirements—popular AI chatbots provided incorrect information 50% of the time.</li> <li>A study by the AI Democracy Projects indicated that “51% of the answers provided by chatbots were inaccurate; 40% were harmful; 38% included incomplete information; and 13% were biased.”<sup>294</sup></li> </ul>	AI-enabled misinformation	<ul style="list-style-type: none"> <li>AI firms responded with criticisms of the study's methodology and noted that their products may perform differently when accessed through an API.</li> </ul>
US	2024	<p><b>Kremlin-Generated Kamala Harris Deepfake Video:</b></p> <ul style="list-style-type: none"> <li>The Kremlin allegedly spread a deepfake video claiming Harris hit a 13-year-old girl, left her paralyzed, and ran.<sup>295</sup></li> </ul>	AI-generated deepfake video	<ul style="list-style-type: none"> <li>Russia supports Donald Trump and is attempting to diminish Harris’ chances of winning the election.</li> </ul>
US	2024	<p><b>Kamala Harris Deepfake Videos:</b></p> <ul style="list-style-type: none"> <li>Elon Musk shared a deepfake video of Vice President Kamala Harris describing herself as a “diversity hire,” in relation to her role as the Democratic Party’s choice for the US presidency.</li> <li>The post violates X’s policy on synthetic video as it did not include a statement disclosing the video as a parody.</li> </ul>	AI-generated deepfake video	<ul style="list-style-type: none"> <li>Lawmaker—including Senator Amy Klobuchar, Representative Barbara Lee, and Governor Gavin Newsom—criticized the video and called for tougher regulation on content produced by AI.<sup>296</sup></li> </ul>
US	2024	<p><b>Foreign adversaries using ChatGPT to influence US elections</b><sup>297</sup></p> <ul style="list-style-type: none"> <li>Iran, North Korea, Russia, and China using AI to influence US presidential elections.</li> <li>The ODNI mentioned that governments like Russia and Iran are changing their influence operations strategy.<sup>298</sup></li> </ul>	AI-enabled influence operations	<ul style="list-style-type: none"> <li>No significant impact yet.</li> <li>Iran likely to pose as activists that support pro-Gaza protests<sup>299</sup> to exacerbate domestic divisions in the future.<sup>300</sup></li> <li>In August 2024, OpenAI discovered and disrupted Iranian groups using ChatGPT to generate social media posts that increase political division. However, the content did not receive high engagement.<sup>301</sup></li> </ul>

Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
US	2024	<p><b>Jailbreaking ChatGPT:</b></p> <ul style="list-style-type: none"> <li>Threat actors are using “Skeleton Key” attacks to jailbreak OpenAI’s ChatGPT and force it to generate inappropriate responses. <sup>302</sup></li> </ul>	AI jailbreak	<ul style="list-style-type: none"> <li>ChatGPT could be then utilized to write disinformation, share sensitive information, etc. which can interfere with the US election.</li> </ul>
US	2024	<p><b>AsyncRAT Malware:</b></p> <ul style="list-style-type: none"> <li>HP researchers discovered a malicious campaign targeting French speakers with AsyncRAT.</li> <li>The malware records victims’ screens and keystrokes.</li> <li>The structure of the malware appears to suggest that the code was generated by AI. <sup>303</sup></li> </ul>	AI-generated malware/code	<ul style="list-style-type: none"> <li>Enhanced microtargeting of a specific community in the US ahead of the presidential election.</li> </ul>
US	2024	<p><b>Chinese Influence Operation “Spamouflage”:</b><sup>304</sup></p> <ul style="list-style-type: none"> <li>In August 2024, Meta removed nearly 9,000 accounts linked to Spamouflage.</li> <li>Used AI-generated content to sow division in the US, especially related to the War in Gaza.</li> <li>Created fake personas of voters across social media platforms. <sup>305</sup></li> </ul>	AI-generated content, AI-generated fake personas	<ul style="list-style-type: none"> <li>Pro-Chinese rhetoric trying to support the Republican party.</li> <li>Other influence operations used Spamouflage to gain an audience.</li> <li>Pro-China account Harlan Report posed as a US conservative media outlet and influencer promoting Trump. IT gained an online following from Spamouflage’s support.</li> <li>Harlan Report mocked and criticized President Biden and constantly switched identities. <sup>306</sup></li> </ul>
Slovakia	2024	<p><b>Russian Interference in Slovakian Elections:</b><sup>307</sup></p> <ul style="list-style-type: none"> <li>Deepfake audios of Progressive Slovakia’s candidate Michal Šimečka and Monika Tódová discussing election rigging emerged ahead of elections.</li> <li>The audio, spread during a pre-election moratorium, was likely linked to the Kremlin and amplified by Russian state media.</li> </ul>	AI-Generated Deepfake	<ul style="list-style-type: none"> <li>Progressive Slovakia lost to SMER.</li> <li>SMER campaigned against military support for Ukraine.</li> </ul>

Country	Date	Description	AI Use	Impact, Concerns, and Suggestions
US/UK	2024	<p><b>Russian Interference in US Elections and UK Parliamentary Elections:</b><sup>308</sup></p> <ul style="list-style-type: none"> <li>• Proxy-media websites questioned US democracy and promoted conspiracy theories with content appearing to be altered from mainstream US news outlets.</li> <li>• The campaign targeted US and European elections and included themes on politics, migration, and border security in France and Germany.</li> </ul>	AI-Generated Fake Media Stories	<ul style="list-style-type: none"> <li>• Though these webpages appear to still publish conspiracy theories and other false information, major social media companies like Meta have worked to remove content from their networks and have reported that the removal occurred before major authentic audience engagement was gained.</li> </ul>
Moldova	2023	<p><b>Russian Interference in Moldovan Election</b><sup>309</sup></p> <ul style="list-style-type: none"> <li>• A deepfake video emerged of Moldovan President Maia Sandu.</li> <li>• It falsely depicted Sandu speaking negatively about Moldovans and calling George Soros and the US the “sponsors” of her administration.</li> <li>• The video emerged on Russian-language channels, aiming to influence public opinion against Sandu.</li> </ul>	AI-Generated Deepfake	<ul style="list-style-type: none"> <li>• The president denied the authenticity of the statements made in the deepfake video.</li> <li>• Major social media companies, including Meta and TikTok have expressed a commitment to responding to the country’s reports of disinformation.</li> </ul>
Pakistan <sup>310</sup>	2024	<p><b>Interference in Domestic Elections</b></p> <ul style="list-style-type: none"> <li>• Deepfake audio of Imran Khan suggesting an election boycott and a deepfake video of political party Pakistan Tehreek-e-Insaf (PTI)-backed independent candidate Raja Bashara renouncing politics.</li> <li>• These deepfakes aimed to confuse PTI supporters and reduce electoral support.</li> </ul>	AI-Generated Deepfake Audio	<ul style="list-style-type: none"> <li>• The PTI social media team claimed the video was constructed based on notes from Imran Khan while in prison, but activists expressed concern that even though disclaimers were present on the video, it is difficult for voters to understand if the comments in the deepfake originated from Khan or the social media team.</li> </ul>
Rwanda <sup>311</sup>	2024	<ul style="list-style-type: none"> <li>• ChatGPT was used to generate social media content about the election.</li> </ul>	AI-Generated Posts	<ul style="list-style-type: none"> <li>• The posts did not get many likes, views, or shares.</li> </ul>

## Appendix B: Overview of Recent Global AI Regulation<sup>312</sup>

Authority and Regulation	Implementation Date	Description	Policymaking Approach
European Union (EU) proposed AI Act	Early 2025	<ul style="list-style-type: none"> <li>Stringent rules governing high-risk AI systems, transparency, and data governance measures.</li> <li>The latest draft bans the bulk scraping of facial images to build databases, social scoring, and emotion recognition in the workplace.</li> <li>Financial penalties for non-compliance, of up to 7% of annual global revenues.</li> <li>Though the act's jurisdiction is limited to the EU, it will have extraterritorial impacts given its applicability to all entities with operations in the EU.</li> <li>Before the law comes into effect, the EU is asking companies to voluntarily commit to adhering to key parts of the Act by signing an AI Pact.</li> </ul>	Risk- and Rules-Based
United Nations Global Digital Compact Process	September 2024	<ul style="list-style-type: none"> <li>Established a High-Level Advisory Body for AI to gather experts from states, UN entities, industry, academia, and civil society.</li> <li>Seeks to provide recommendations on international AI governance.</li> <li>Includes a digital human rights advisory mechanism facilitated by the Office of the High Commissioner for Human Rights (OHCHR), designed to offer practical guidance on the intersection of human rights and technology issues.</li> </ul>	Principles-based
Singapore The Model AI Governance Framework	May 2024	<ul style="list-style-type: none"> <li>Stipulates nine dimensions: Accountability, Data, Trusted Development and Deployment, Incident Reporting, Testing and Assurance, Security, Content Provenance, Safety and Alignment R&amp;D and AI for Public Good.</li> </ul>	Principles-based
G7 Hiroshima Process	October 2023	<ul style="list-style-type: none"> <li>The Hiroshima AI Process Comprehensive Policy Framework was established, including guiding principles and code of conduct aimed at promoting safe, secure, and trustworthy AI systems.</li> </ul>	Principles-based
UK Bletchley Declaration	November 2023	<ul style="list-style-type: none"> <li>Encourages organizations to practice context-appropriate transparency and accountability in measuring, monitoring, and mitigating potentially harmful AI capabilities and their effects.</li> <li>Focuses on preventing misuse, control issues, and the amplification of other risks.</li> <li>Aims to develop a shared understanding of AI risks.</li> <li>Promotes the creation of risk-based policies, including transparency requirements, evaluation metrics, safety testing tools, and public sector scientific research.</li> </ul>	Risk-based

Authority and Regulation	Implementation Date	Description	Policymaking Approach
China New/Next Generation and Submission Artificial Intelligence Development Plan and Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment)	2017	<ul style="list-style-type: none"> <li>• Aims for global leadership in AI research by 2030.</li> <li>• Promotes AI in sectors like manufacturing and healthcare.</li> <li>• Seeks to develop a robust ethical AI governance framework.</li> <li>• Prioritizes AI civil-military integration before AI infrastructure.</li> <li>• Recognizes the need for AI safety regulation.</li> <li>• Enforce content restrictions to align with Socialist Core Values.</li> <li>• Prohibit content undermining state power, inciting separatism, or disturbing social order.</li> <li>• Reflect political priorities and concerns about AI's impact on stability and order.</li> </ul>	Rules-based

## Appendix C: Comparison Between the NIST AI RMF and Japan’s AI GfB

Aspect	Similarities	Differences
Focus on Trustworthiness & Risk Management	Both frameworks emphasize AI systems must be trustworthy, focusing on safety, security, transparency, and fairness.	The AI RMF provides detailed definitions and criteria, while AI GfB integrates these into broader guiding principles.
Safety Considerations	Both stress ensuring AI systems do not harm human health, property, or environment.	
Transparency & Accountability	Both advocate for transparency, requiring relevant information to be available and emphasizing accountability.	NIST defines accountability and transparency explicitly. AI GfB addresses these concepts more broadly.
Performance and Reliability	Both highlight the need for AI systems to perform reliably under defined conditions and maintain performance.	
Definitions & Terminology	NIST provides detailed definitions for validation, reliability, accuracy, robustness.	AI GfB lacks specific definitions for some terms.
Security & Resilience	NIST distinguishes between security (confidentiality, integrity, availability) and resilience (withstanding adverse events).	AI GfB focuses on general security measures and acknowledges that vulnerabilities cannot be eliminated.
Bias & Fairness	NIST categorizes biases into systemic, computational, and human-cognitive.	AI GfB addresses fairness and bias broadly without specific categories.
Explainability & Interpretability	NIST defines explainability as understanding AI mechanisms and interpretability as understanding AI outputs.	AI GfB includes these concepts under transparency and accountability without specific definitions.
Privacy Considerations	NIST emphasizes norms and practices to safeguard personal data and autonomy.	AI GfB references privacy protection generally but lacks a specific definition.
Human-Centric Approach	NIST focuses on upholding human rights and respect in AI systems.	AI GfB explicitly incorporates a human-centric approach, emphasizing human dignity and ethical education.
Resilience	NIST defines resilience as maintaining functionality and adapting to changes.	AI GfB does not define resilience but focuses more on security and acknowledges existing vulnerabilities.
Context & Application	NIST provides detailed guidance on implementing trustworthiness across various AI lifecycle stages.	AI GfB focuses on broader guiding principles, emphasizing a human-centric approach and societal impacts.

## Appendix D: Table of State-Level AI Regulation

State	Legislation	Date	Description	Approach
California	Senate Bill 1047	February 2024	<ul style="list-style-type: none"> <li>• If passed, the legislation would mandate strict compliance requirements for large or powerful AI models that are theoretically capable of certain harmful capabilities.</li> <li>• Developers could face civil and even criminal liability for any violation of these mandates.</li> <li>• 30 new measures on AI aimed at protecting consumers and jobs, deemed one of the biggest efforts of regulation.</li> <li>• Include rules to prevent AI tools from discriminating in housing and health care services, protect IP and jobs.</li> <li>• Mandates “kill switch” for applicable AI models.<sup>313</sup></li> <li>• Requires developers to integrate safeguards as they develop and deploy what the bill calls “covered models.”</li> <li>• Many AI companies oppose the bill.<sup>314</sup></li> <li>• In late September 2024, California Governor Newsom vetoed the bill due to concerns that the bill narrowly focused on large models, potentially leaving smaller but equally risky AI systems unchecked and imposing excessive regulatory burdens that could hinder innovation.</li> <li>• The veto further sparked <a href="#">debate on the balance between AI safety and innovation</a>.</li> <li>• Supporters argue the decision delays crucial safeguards in a rapidly advancing technology field, while critics believe it avoids restrictive measures that might stunt AI growth.<sup>315</sup></li> </ul>	Rules-based
	California Consumer Privacy Act <sup>316</sup>	2018	<ul style="list-style-type: none"> <li>• Contains provisions on the use of automated decision-making tools (ADMT).</li> <li>• California Privacy Protection Agency released draft rules on these provisions governing consumer notice, access, and opt-out rights concerning automated decision-making technology.</li> <li>• Regulations are still being finalized but will likely cover expanded uses of AI.</li> <li>• Draft rules – expected to be formalized in 2024 – would require significant disclosure about businesses’ implementation and use of ADMT.</li> </ul>	Risk-based
		2020	<ul style="list-style-type: none"> <li>• California voters approved Proposition 24, the CPRA, which amended the CCPA and added new additional privacy protections beginning in 2023.</li> </ul>	Rules-based
Colorado	CO BS 113 <sup>317</sup>	2020	<ul style="list-style-type: none"> <li>• Establishes provisions for government use of facial recognition technology.</li> <li>• Requires state and local agencies intending to use facial recognition technology to file a report stating their intent to develop or procure facial recognition technology and specify how they will use facial recognition.<sup>318</sup></li> </ul>	Risk-based



State	Legislation	Date	Description	Approach
			<ul style="list-style-type: none"> <li>Requires agencies to develop a data management policy, establish testing procedures, and provide information on false identifications.</li> </ul>	
	CO BS 205 <sup>319</sup>	2024	<ul style="list-style-type: none"> <li>Deployers must notify consumers when “high-risk” AI systems influence consequential decisions.</li> <li>Provide disclosures about AI usage.</li> <li>Offers opt-out option for personal data processing.</li> <li>Allow corrections to data and appeals.</li> <li>Mandate human review of decisions.</li> <li>Exemption for those with less than 50 employees.</li> <li>Developers of AI systems that interact with consumers must disclose that the interaction is with an AI system.</li> </ul>	Rules-based
Michigan	Series of bills aimed at addressing deepfakes in elections <sup>320</sup>	2023	<ul style="list-style-type: none"> <li>MI HB 5141, which requires disclosure for pre-recorded phone messages or political advertisements generated with AI.</li> <li>MI HB 5143, which defines “artificial intelligence” as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.</li> <li>MI HB 5144, which prohibits distributing media that manipulates the speech or conduct meant to deceive voters within 90 days of an election unless a disclaimer is provided.</li> <li>MI HB 5145, which establishes sentencing guidelines for election law offenses involving materially deceptive media.</li> </ul>	Rules-based
New York	NY AB 8808 & SB 8308 <sup>321</sup>	2024	<ul style="list-style-type: none"> <li>As part of a broader budget law, lawmakers enacted a right of action against non-consensual sexual deepfakes and a prohibition on political deepfakes without a disclaimer.</li> </ul>	Rules-based
	NY SB 1042A <sup>322</sup>	2023	<ul style="list-style-type: none"> <li>Makes it a crime to intentionally disseminate or publish deepfake content that depicts someone with "one or more intimate parts exposed or engaging in sexual conduct with another person, including an image created or altered by digitization, where such person may reasonably be identified."</li> </ul>	Rules-based
	Local Law 144 <sup>323</sup>	2021 <sup>324</sup>	<ul style="list-style-type: none"> <li>Prohibits employers and employment agencies from using an automated employment decision tool (AEDT) in New York City unless they ensure a bias audit was done and provide required notices.</li> </ul>	Risk-based
Florida	FL HB 919 <sup>325</sup>	2024	<ul style="list-style-type: none"> <li>Requires political advertising to include a disclaimer if it uses GenAI that appears to depict a real person performing an action that did not occur and was created with the intent to injure a candidate or to deceive regarding a ballot issue.</li> <li>Law provides for civil and criminal penalties.</li> </ul>	Rules-based
	FL SB 1680 <sup>326</sup>	2024	<ul style="list-style-type: none"> <li>Law created the Florida Government Technology Modernization Council to study and monitor the development and deployment of AI systems.</li> </ul>	Principles-based



## Appendix E: Strengths and Weaknesses of Select US AI Governance Approach

Framework	Strengths	Weaknesses
<b>EO14110</b>	<ul style="list-style-type: none"> <li>• Comprehensive scope: address a wide range of AI governance issues</li> <li>• Clear directives for federal agencies and private enterprises</li> <li>• Promotion of innovation through support for AI research, technical assistance, and streamlined visa procedures for AI talent</li> <li>• Privacy protection focusing on protecting consumer privacy through guidelines and promoting privacy-enhancing technologies (PETs)</li> <li>• Focus on fairness, equity, and privacy to address societal concerns and enhance public trust in AI</li> <li>• Aims to strengthen US leadership in AI globally through standardization and cooperation</li> <li>• Structured approach to regulating AI across various sectors</li> <li>• Integration of AI within federal agencies, enhancing operational efficiency and service delivery</li> </ul>	<ul style="list-style-type: none"> <li>• Broad scope may lead to challenges in consistent implementation across different sectors/agencies</li> <li>• Compliance requirements (reporting, cybersecurity standards) may be resource-intensive for smaller organizations</li> <li>• Potential overreach in regulating vs fostering innovation</li> <li>• Privacy guidelines may not keep pace with rapidly evolving AI technologies</li> </ul>
<b>AI Bill of Rights</b>	<ul style="list-style-type: none"> <li>• Provides sector-specific guidance with actions that address most high-priority algorithmic harms across healthcare, financial services, education, and housing</li> <li>• Foundation to build capacity as it covers a wide range of issues and federal actions</li> <li>• Importance of having international AI regulation</li> </ul>	<ul style="list-style-type: none"> <li>• Nonbinding principles fail to keep individuals and organizations that do not adhere to the Bill of Rights accountable</li> <li>• No coordination with the AI EO or other proposed AI governance approaches</li> <li>• Uneven progress between sectors leaves sectors like education and workplace surveillance behind compared to the health and finance sectors</li> </ul>
<b>NIST RMF</b>	<ul style="list-style-type: none"> <li>• Flexible and applicable across use cases</li> <li>• Recognizes that risk is context-dependent</li> <li>• Builds on existing knowledge around privacy and security, rather than approaching AI as totally novel</li> <li>• Consideration of societal impacts and human behavior in AI decision-making processes</li> <li>• Alignment with global standards, facilitating global cooperation</li> <li>• Importance of public input and iterative development</li> <li>• Multi-stakeholder approach has developed goodwill from the industry</li> <li>• Early mover advantage that could lead to increased uptake and influence in understanding how to ensure trustworthy AI in practice</li> </ul>	<ul style="list-style-type: none"> <li>• Delves deeply into technical details which may be challenging for individuals outside the field to understand</li> <li>• Less effective for stakeholders in the policy domain who lack a strong background in technology</li> <li>• Resource intensive</li> </ul>

Framework	Strengths	Weaknesses
<b>NIST SSDF</b>	<ul style="list-style-type: none"> <li>• Adapts preexisting software development practices for AI governance</li> <li>• Covers various phases of the AI lifecycle to mitigate harms at every phase</li> <li>• Aligns the technical pieces with the business mission and organizational goals, incorporating AI governance into businesses more effectively</li> <li>• Aims to increase trust and transparency, enhancing the credibility of AI-generated content, as well as contributing to AI digital literacy.</li> </ul>	<ul style="list-style-type: none"> <li>• One framework provided across all industries which may not be functional in certain sectors</li> <li>• Difficult to implement due to the lack of specific actions and guidance</li> <li>• Not easily accessible and understood by non-technical teams and individuals</li> </ul>
<b>California: CA SB-1047<sup>327</sup></b>	<ul style="list-style-type: none"> <li>• New regulatory framework specifically for advanced AI systems, setting a threshold based on computational capacity (1026 FLOP).</li> <li>• Establishes a regulatory precedent for AI governance, potentially influencing federal policy</li> <li>• Addressing foreseeable risks associated with advanced AI systems</li> <li>• Offers two compliance pathways for developers: a limited duty exemption and implementation of specified safeguards.</li> <li>• Incentives for compliance: preventative measures (restraining orders, injunctions) for imminent threats, compensatory and punitive damages for actual harm caused</li> <li>• Focus on public safety: targets AI systems capable of causing significant harm, e.g. facilitating cyberattacks or weapons development, ensuring that regulatory attention is on high-risk applications</li> </ul>	<ul style="list-style-type: none"> <li>• Limited scope of liability, primarily tied to failure to adopt specific precautionary measures rather than strict liability for all harms caused by the systems.</li> <li>• Bill acknowledges that current safety measures may not be sufficient</li> <li>• Compliance pathways and the subjective assessment of “reasonable assessment” may lead to varying interpretations and enforcement</li> <li>• Concerns that stringent regulations may hinder open-source AI development, which could otherwise contribute to innovation</li> </ul>

## 11. Annotated Bibliography

The below annotated bibliography presents an overview of the ten key sources used in this report's literature review as well as their contribution to the report's analysis and key findings.

**1. NIST, Four Principles of Explainable Artificial Intelligence. 2021. <https://doi.org/10.6028/NIST.IR.8312>**

This report introduces four key principles for creating explainable AI systems: (1) Explanation, (2) Meaningfulness, (3) Explanation Accuracy, and (4) Knowledge Limits. These principles aim to ensure that AI systems provide understandable and accurate explanations that help users trust and utilize AI effectively. These four principles were developed to encompass the multidisciplinary nature of explainable AI, including the fields of computer science, engineering, and psychology.

**2. NIST, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models. April 2024.**

**<https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>**

This publication from NIST presents the Secure Software Development Framework (SSDF), which provides best practices for generative AI and recommendations for reducing vulnerabilities in software development. The SSDF is intended to help organizations integrate security practices into their software development lifecycle (SDLC), offering actionable guidance to address risks and prevent vulnerabilities in software systems.

**3. The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>**

This executive order from the Biden administration calls for the safe, secure, and trustworthy development and use of AI. The order emphasizes the importance of safeguarding civil rights, privacy, and national security while promoting the ethical use of AI technology. Key actions in the order include directives for federal agencies to establish frameworks that monitor AI systems' performance, especially in critical sectors such as healthcare, education, and law enforcement. The executive order mandates that AI systems used by federal agencies meet strict guidelines for transparency, accuracy, and fairness.

**4. The White House Office of Science and Technology Policy, Blueprint for an AI Bill of Rights, 2022. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>**

This document provides guidance on protecting the public from the risks posed by AI technologies and outlines five key principles designed to ensure the safe, equitable, and transparent development and use of AI technologies. These principles are: (1) Safe and Effective Systems, (2) Algorithmic Discrimination Protections, (3) Data Privacy, (4) Notice and Explanation, and (5) Human Alternatives, Consideration, and Fallback.

**5. Carlson, John. *Treasury: AI-fueled Cyber Threats Bring New Challenges*. ABA Banking Journal, April 11, 2024. <https://bankingjournal.aba.com/2024/04/treasury-ai-fueled-cyber-threats-bring-new-challenges/>**

This article discusses a report from the U.S. Department of the Treasury on the rising cyber threats powered by AI and the challenges they pose to financial institutions. The Treasury warns that AI-driven cyberattacks, such as automated phishing and malware, are becoming more sophisticated, making it difficult for traditional security measures to keep up. The report highlights how malicious actors are leveraging AI to exploit vulnerabilities in banking systems and to target sensitive financial data more effectively. It also emphasizes the need for financial institutions to adopt AI-enhanced cybersecurity tools and frameworks to mitigate these emerging risks.

**6. Goldston, David, Huttenlocher, Dan, Ozdaglar, Asu. A Framework for U.S. AI Governance: Creating a Safe and Thriving AI Sector, November 28, 2023.**

<https://computing.mit.edu/wp-content/uploads/2023/11/AIPolicyBrief.pdf>

This framework discusses the ethical implications and governance challenges of AI. It emphasizes the need for frameworks that ensure AI development aligns with societal values, promoting transparency, accountability, and inclusivity. The brief calls for collaborative efforts among stakeholders—governments, industries, and civil society—to create policies that balance innovation with ethical standards, aiming to protect human rights while leveraging AI's benefits.

**7. World Economic Forum. AI Governance Alliance Briefing Paper Series. World Economic Forum, January 2024. [https://www3.weforum.org/docs/WEF\\_AI\\_Governance\\_Alliance\\_Briefing\\_Paper\\_Series\\_2024.pdf](https://www3.weforum.org/docs/WEF_AI_Governance_Alliance_Briefing_Paper_Series_2024.pdf)**

This briefing is comprised of three briefing papers that outline the critical challenges and opportunities in AI governance. It outlines principles for ensuring trustworthy AI, including accountability, transparency, and human-centric values. The paper advocates for multi-stakeholder collaboration to address the ethical, legal, and social implications of AI technologies. It calls for the establishment of international standards and guidelines to navigate the rapidly evolving AI landscape, ensuring that innovations benefit society while mitigating potential risks.

**8. Aspen Institute. *Generative AI Regulation and Cybersecurity*. Aspen Digital, 2024.**

[https://www.aspeninstitute.org/wp-content/uploads/2024/03/Aspen-Digital\\_Generative-AI-Regulation-and-Cybersecurity\\_January-2024.pdf](https://www.aspeninstitute.org/wp-content/uploads/2024/03/Aspen-Digital_Generative-AI-Regulation-and-Cybersecurity_January-2024.pdf)

This report discusses the dual role of generative AI in cybersecurity, presenting both risks and opportunities. It calls for balanced regulation to ensure ethical development while encouraging innovation. The report emphasizes collaboration between governments and industries, promoting transparency and human oversight to mitigate risks. The report also notes details how as generative AI technology rapidly evolves, effective governance structures are crucial to address the challenges and maximize its benefits.

**9. UNESCO, Recommendation on the Ethics of Artificial Intelligence, November 2021.**

<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

This report by UNESCO presents a comprehensive set of recommendations aimed at guiding the ethical development and deployment of AI technologies. It highlights the importance of aligning AI systems with human rights and fundamental freedoms, ensuring that they serve the public good and promote social justice. The recommendation covers various ethical principles, including transparency, accountability, fairness, and the need for inclusive and participatory approaches in AI governance.

**10. Carnegie Endowment for International Peace, *Advancing a More Global Agenda for Trustworthy Artificial Intelligence*. Carnegie Endowment, April 30, 2024.**

<https://carnegieendowment.org/research/2024/04/advancing-a-more-global-agenda-for-trustworthy-artificial-intelligence?lang=en>

This report examines the challenges of AI development and governance, particularly its impact on the "Global Majority" - communities in the Global South and marginalized groups worldwide. The report discusses issues such as data divides, underrepresentation of Global Majority languages in AI training, and the potential for AI systems to undermine trust when they don't consider local contexts. The report calls for more inclusive, context-aware AI governance to ensure equitable AI development and deployment on a global scale.

## 12. References

1. "A Framework for Election Vendor Oversight | Brennan Center for Justice." Accessed September 23, 2024. <https://www.brennancenter.org/our-work/policy-solutions/framework-election-vendor-oversight>.
2. Access Partnership. "Brazil's New AI Bill: A Comprehensive Framework for Ethical and Responsible Use of AI Systems," May 5, 2023. <https://accesspartnership.com/access-alert-brazils-new-ai-bill-a-comprehensive-framework-for-ethical-and-responsible-use-of-ai-systems/>.
3. "AI and Elections: Lessons for Southeast Asia - RSIS." Accessed September 23, 2024. <https://www.rsis.edu.sg/rsis-publication/fit/ai-and-elections-lessons-for-southeast-asia/>.
4. "AI and India's General Elections." Accessed September 23, 2024. <https://thediplomat.com/2024/04/ai-and-indias-general-elections/>.
5. "AI and India's General Elections – The Diplomat." Accessed September 23, 2024. <https://thediplomat.com/2024/04/ai-and-indias-general-elections/>.
6. "AI Policy Overview: Michigan." Accessed August 31, 2024. <https://www.multistate.ai/ai-policy-overview-michigan>.
7. "AI Will Increase the Quantity — and Quality — of Phishing Scams." Accessed September 23, 2024. <https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams>.
8. AI.gov. "Federal AI Use Case Inventories." Accessed August 29, 2024. <https://ai.gov/ai-use-cases/>.
9. Alliance for Justice. "AI Threats in Elections: What Nonprofits Must Know," July 16, 2024. <https://afj.org/article/ai-threats-in-elections-what-nonprofits-must-know/>.
10. Angwin, Julia, Alondra Nelson, and Rina Palta. "Seeking Reliable Election Information? Don't Trust AI," n.d.
11. Arnold & Porter. "Department of Commerce Proposes Rule on IaaS Product-Related Customer Identification and AI-Related Reporting Requirements | Advisories," January 30, 2024. <https://www.arnoldporter.com/en/perspectives/advisories/2024/01/dept-of-commerce-proposes-rule-on-iaas-product-related>.
12. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." NIST AI 600-1 Initial Public Draft. National Institute of Standards and Technology, April 2024. <https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf>.
13. Ash Center. "AI and the 2024 Elections," May 29, 2024. <https://ash.harvard.edu/articles/ai-and-the-2024-elections/>.
14. Aspen Digital. "Generative AI Regulation and Cybersecurity." Aspen Institute, 2024. <https://www.aspendigital.org/report/generative-ai-regulation-and-cybersecurity/>.
15. Aspen Digital. "Generative AI Regulation and Cybersecurity." Accessed September 23, 2024. <https://www.aspendigital.org/report/generative-ai-regulation-and-cybersecurity/>.
16. Baker, Tessa. "What Does AI Red-Teaming Actually Mean?" *Center for Security and Emerging Technology* (blog), October 24, 2023. <https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/>.
17. Ball, Molly. "How COVID-19 Changed Everything About the 2020 Election." TIME, August 6, 2020. <https://time.com/5876599/election-2020-coronavirus/>.
18. "Biden Deepfake Spreads Online after Withdrawal from 2024 Race | Fact Check." Accessed September 23, 2024. <https://factcheck.afp.com/doc.afp.com.364R2N9>.
19. Booth, Harold, Murugiah Souppaya, Apostol Vassilev, Michael Ogata, Martin Stanley, and Karen Scarfone. "Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile." National Institute of Standards and Technology, April 29, 2024. <https://doi.org/10.6028/NIST.SP.800-218A.ipd>.
20. Brennan Center for Justice. "Artificial Intelligence Legislation Tracker," August 2024. <https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-legislation-tracker>.

21. Brennan Center. "Safeguards for Using Artificial Intelligence in Election Administration | Brennan Center for Justice," November 3, 2023. <https://www.brennancenter.org/our-work/research-reports/safeguards-using-artificial-intelligence-election-administration>.
22. Brookings. "The Impact of Generative AI in a Global Election Year." Accessed September 23, 2024. <https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/>.
23. Brookings. "What Role Is AI Playing in Election Disinformation?" Accessed September 23, 2024. <https://www.brookings.edu/articles/what-role-is-ai-playing-in-election-disinformation/>.
24. Bryan Cave Leighton Paisner LLP. "US State-by-State AI Legislation Snapshot." BCLP - Bryan Cave Leighton Paisner - US state-by-state AI legislation snapshot, 2024. <https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html>.
25. "California Proposes 30 AI Regulation Laws Amid Federal Standstill - The New York Times." Accessed June 23, 2024. <https://www.nytimes.com/2024/06/10/technology/california-ai-regulation.html>.
26. Campbell, James. "Influence Actors Likely to Adjust Tactics Amid Election Chaos." *The Record*, September 24, 2024. <https://therecord.media/influence-actors-likely-to-adjust-tactics-amid-election-chaos>.
27. Campbell, James. "Spamouflage: The Influence Operation Behind China's Information Manipulation." *The Record*, September 25, 2024. <https://therecord.media/spamouflage-influence-operation-china>.
28. Carnegie Endowment for International Peace. "California SB 1047: AI Safety Bill Veto Lessons." Last modified October 2024. Accessed October 29, 2024. <https://carnegieendowment.org/posts/2024/10/california-sb1047-ai-safety-bill-veto-lessons?lang=en>.
29. CDO Magazine Bureau. "US Department of Education Releases Guidelines for Integrating AI into Edtech." CDO Magazine, July 12, 2024. <https://www.cdomagazine.tech/us-federal-news-bureau/us-department-of-education-releases-guidelines-for-integrating-ai-into-edtech>.
30. Center for Internet Security. "Election Infrastructure Information Sharing and Analysis Center (EI-ISAC)." Accessed September 26, 2024. <https://www.cisecurity.org/ei-isac>.
31. Centre For Public Impact (CPI). "The Good, the Bad and the Ugly Uses of Machine Learning in Election Campaigns." Accessed September 23, 2024. <https://www.centreforpublicimpact.org/insights/good-bad-ugly-uses-machine-learning-election-campaigns>.
32. "ChatGPT Creates Mutating Malware That Evades Detection by EDR | CSO Online." Accessed September 23, 2024. <https://www.csoonline.com/article/575487/chatgpt-creates-mutating-malware-that-evades-detection-by-edr.html>.
33. Chia, Austin. "AI Bill of Rights: What Does It Mean?" Splunk, July 24, 2024. [https://www.splunk.com/en\\_us/blog/learn/ai-bill-of-rights.html](https://www.splunk.com/en_us/blog/learn/ai-bill-of-rights.html).
34. CISOMAG. "Artificial Intelligence as Security Solution and Weaponization by Hackers." *CISO MAG | Cyber Security Magazine* (blog), December 9, 2019. <https://cisomag.com/hackers-using-ai/>.
35. Civil Code - CIV DIVISION 3. OBLIGATIONS [1427 - 3273.69] ( Heading of Division 3 amended by Stats. 1988, Ch. 160, Sec. 14. ) PART 4. OBLIGATIONS ARISING FROM PARTICULAR TRANSACTIONS [1738 - 3273.69] ( Part 4 enacted 1872. ). Accessed August 31, 2024. [https://leginfo.legislature.ca.gov/faces/codes\\_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5](https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5).
36. "Clinton Wins Popular Vote by Nearly 2.9 Million | AP News." Accessed September 23, 2024. <https://apnews.com/article/2c7a5afc13824161a25d8574e10ff4e7>.
37. Cohen, Richard. "Artificial Intelligence Regulations: Biden and Trump Need to Act." *Chicago Tribune*, June 10, 2024. <https://www.chicagotribune.com/2024/06/10/opinion-artificial-intelligence-ai-joe-biden-donald-trump-regulations/>.
38. Computer Security Division, Information Technology Laboratory. "NIST SSDF for Generative AI and Dual Use Foundation Models | CSRC." CSRC | NIST, January 8, 2024. <https://csrc.nist.gov/Events/2024/nist-ssdf-for-generative-ai-dual-use-foundation>.



39. "Conspiracy Theories That Biden Is Dead, Dying Boosted by Musk's X." Accessed September 23, 2024. <https://www.nbcnews.com/tech/internet/joe-biden-dead-dying-covid-vegas-theory-musk-x-elon-rcna163248>.
40. Consumer Financial Protection Bureau. "CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms," May 26, 2022. <https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/>.
41. Cybersecurity & Infrastructure Security Agency (CISA). "CISA, JCDC, Government and Industry Partners Conduct AI Tabletop Exercise," June 14, 2024. <https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-conduct-ai-tabletop-exercise>.
42. Cybersecurity & Infrastructure Security Agency (CISA). "CISA Roadmap for Artificial Intelligence," November 2023. [https://www.cisa.gov/sites/default/files/2023-11/2023-2024\\_CISA-Roadmap-for-AI\\_508c.pdf](https://www.cisa.gov/sites/default/files/2023-11/2023-2024_CISA-Roadmap-for-AI_508c.pdf).
43. Cybersecurity & Infrastructure Security Agency (CISA). "Election Security." September 27, 2024. <https://www.cisa.gov/topics/election-security>.
44. Cybersecurity & Infrastructure Security Agency. "Government Coordinating Councils." Accessed September 26, 2024. <https://www.cisa.gov/resources-tools/groups/government-coordinating-councils>.
45. Cybersecurity & Infrastructure Security Agency. *Physical Security Checklist for Election Offices*. September 2024. <https://www.cisa.gov/sites/default/files/2024-09/Physical-Security-Checklist-for-Election-Offices-508.pdf>.
46. Cybersecurity & Infrastructure Security Agency. *Readiness and Resilience Checklist for Election Offices*. September 2024. <https://www.cisa.gov/sites/default/files/2024-09/Readiness-and-Resilience-Checklist-for-Election-Offices-508.pdf>.
47. Cybersecurity & Infrastructure Security Agency (CISA). "Risk in Focus: Generative A.I. and the 2024 Election Cycle," 2024. <https://www.cisa.gov/resources-tools/resources/risk-focus-generative-ai-and-2024-election-cycle>
48. Cybersecurity & Infrastructure Security Agency (CISA). "Sector Coordinating Councils." Accessed September 26, 2024. <https://www.cisa.gov/resources-tools/groups/sector-coordinating-councils>.
49. Cybersecurity & Infrastructure Security Agency (CISA). "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024. [https://www.cisa.gov/sites/default/files/2024-04/Securing\\_Election\\_Infrastructure\\_Against\\_the\\_Tactics\\_of\\_Foreign\\_Malign\\_Influence\\_Operations\\_2024FINAL\\_508c.pdf](https://www.cisa.gov/sites/default/files/2024-04/Securing_Election_Infrastructure_Against_the_Tactics_of_Foreign_Malign_Influence_Operations_2024FINAL_508c.pdf).
49. Daniels, Owen J., and Dewey Murdick. "Enabling Principles for AI Governance." Center for Security and Emerging Technology, July 2024. <https://cset.georgetown.edu/publication/enabling-principles-for-ai-governance/>.
50. "PSA: Just So You Know: DDoS Attacks Could Hinder Access to Election Information, Would Not Prevent Voting." Accessed November 28, 2024. <https://www.cisa.gov/resources-tools/resources/psa-just-so-you-know-ddos-attacks-could-hinder-access-election-information-would-not-prevent-voting>
51. "Deepfake of Kamala Harris Reups Questions on Tech's Self-Regulation | Council on Foreign Relations." Accessed September 23, 2024. <https://www.cfr.org/blog/deepfake-kamala-harris-reups-questions-techs-self-regulation>.
52. "Deepfakes, Elections, and Shrinking the Liar's Dividend | Brennan Center for Justice," February 8, 2024. <https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend>.
53. Department of Homeland Security (DHS). "DHS Publishes Guidelines and Report to Secure Critical Infrastructure and Weapons of Mass Destruction from AI-Related Threats," April 29, 2024. <https://www.dhs.gov/news/2024/04/29/dhs-publishes-guidelines-and-report-secure-critical-infrastructure-and-weapons-mass>.
54. Department of Homeland Security (DHS). "Fact Sheet: DHS Facilitates Safe and Responsible Deployment and Use of Artificial Intelligence." Last modified April 29, 2024. Accessed October 29, 2024.

- <https://www.dhs.gov/news/2024/04/29/fact-sheet-dhs-facilitates-safe-and-responsible-deployment-and-use-artificial>.
55. Department of Homeland Security (DHS). "Promoting AI Safety and Security," 2024. <https://www.dhs.gov/ai/promoting-ai-safety-and-security>.
  56. DFRLab. "FIMI 101: Understanding Foreign Influence Operations in the Information Environment." September 26, 2024. <https://dfrlab.org/2024/09/26/fimi-101/>.
  57. Dilanian, Ken. "Russia, Iran Using AI to Influence U.S. Election, DNI Warns." *ABC News*, September 25, 2024. <https://abcnews.go.com/Politics/russia-iran-ai-influence-us-election-dni/story?id=113941680>.
  58. "Did Russia Influence Brexit? | Brexit Bits, Bobs, and Blogs | CSIS." Accessed September 23, 2024. <https://www.csis.org/blogs/brexit-bits-bobs-and-blogs/did-russia-influence-brexit>.
  59. DLA Piper. "AI Legislation Advances in U.S. House of Representatives." Last modified October 2024. Accessed October 29, 2024. <https://www.dlapiper.com/en/insights/publications/ai-outlook/2024/ai-legislation-advances-in-us-house-of-representatives>.
  60. Douglas, Karen M., Robbie M. Sutton, and Aleksandra Cichocka. "The Psychology of Conspiracy Theories." *Current Directions in Psychological Science* 26, no. 6 (December 2017): 538–42. <https://doi.org/10.1177/0963721417718261>.
  61. Egan, Matt. "Exclusive: 42% of CEOs Say AI Could Destroy Humanity in Five to Ten Years | CNN Business." CNN, June 14, 2023. <https://www.cnn.com/2023/06/14/business/artificial-intelligence-ceos-warning/index.html>.
  62. "Elections in 2024 and Global Politics." Accessed September 23, 2024. <https://www.reuters.com/graphics/GLOBAL-ELECTIONS2024/gdvzmkejkw/>.
  63. "Elon Musk's AI Fashion Show Goes Viral | Digital Watch Observatory." Accessed September 23, 2024. <https://dig.watch/updates/elon-musks-ai-fashion-show-goes-viral>.
  64. Energy.gov. "DOE AI Risk Management Playbook (AIRMP)." Accessed August 29, 2024. <https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp>.
  65. Engler, Alex. "The AI Bill of Rights Makes Uneven Progress on Algorithmic Protections." Brookings, November 21, 2022. <https://www.brookings.edu/articles/the-ai-bill-of-rights-makes-uneven-progress-on-algorithmic-protections/>.
  66. "Ensuring AI Is Used Responsibly | Homeland Security." Accessed July 24, 2024. <https://www.dhs.gov/ai/ensuring-ai-is-used-responsibly>.
  67. Executive Office of the President. "Maintaining American Leadership in Artificial Intelligence." *Federal Register* 84, no. 30 (February 14, 2019). <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>.
  68. "Exclusive: 42% of CEOs Say AI Could Destroy Humanity in Five to Ten Years | CNN Business." Accessed September 23, 2024. <https://www.cnn.com/2023/06/14/business/artificial-intelligence-ceos-warning/index.html>.
  69. Federal Bureau of Investigation (FBI). "Election Crimes." Accessed November 28, 2024. [Election Crimes — FBI](#)
  70. Firth-Butterfield, Kay, Karen Silverman, and Benjamin Larsen. "Understanding the US 'AI Bill of Rights' - and How It Can Help Keep AI Accountable," October 14, 2022. <https://cdn.jwplayer.com/previews/Gj949ryv-ncRE1zO6>.
  71. "Foreign-Malign-Influence." Accessed September 23, 2024. <https://www.dtra.mil/About/Foreign-Malign-Influence/>.
  72. Franklin, Margarita, Mike Torrey, David Agranovich, and Mike Dvilyanski. "Adversarial Threat Report," n.d.
  73. Freund, Jeffrey. "How We're Ramping Up Our Enforcement of Surveillance Reporting." DOL Blog, September 15, 2022. <http://blog.dol.gov/2022/09/15/how-were-ramping-up-our-enforcement-of-surveillance-reporting>.

74. Garver, Rob. "AI Chatbots Provide False Information About November Elections." Voice of America, March 1, 2024. <https://www.voanews.com/a/ai-chatbots-provide-false-information-about-november-elections/7509355.html>.
75. "Generative AI Poses Threat to Election Security, Federal Intelligence Agencies Warn - CBS News." Accessed September 23, 2024. <https://www.cbsnews.com/news/generative-ai-threat-to-election-security-federal-intelligence-agencies-warn/>.
76. Goldman, Tanya. "What the Blueprint for an AI Bill of Rights Means for Workers." DOL Blog, October 4, 2022. <http://blog.dol.gov/2022/10/04/what-the-blueprint-for-an-ai-bill-of-rights-means-for-workers>.
77. Graphika. "Deepfake It Till You Make It." Accessed September 23, 2024. <https://graphika.com/reports/deepfake-it-till-you-make-it>.
78. Habuka, Hiroki, and David U. Socol de la Osa. "Shaping Global AI Governance: Enhancements and Next Steps for the G7 Hiroshima AI Process," May 24, 2024. <https://www.csis.org/analysis/shaping-global-ai-governance-enhancements-and-next-steps-g7-hiroshima-ai-process>.
79. "Hackers Use AI-Generated Code to Develop Malware, Says HP Threat Report." *The Indian Express*, September 26, 2024. <https://indianexpress.com/article/technology/tech-news-technology/hackers-ai-generated-code-malware-hp-threat-report-9589370/>.
80. Hardy, and Daniel S. Marks. "California Senator Scott Wiener Introduced New Bill That Would Require a Kill Switch for Applicable Artificial Intelligence Models." Benesch Law. Benesch, Friedlander, Coplan & Aronoff LLP - California Senator Scott Wiener Introduced New Bill that Would Require a Kill Switch for Applicable Artificial Intelligence Models, February 15, 2024. <https://www.beneschlaw.com/resources/california-senator-scott-wiener-introduced-new-bill-that-would-require-a-kill-switch-for-applicable-artificial-intelligence-models.html>.
81. Harris, Laurie. "Highlights of the 2023 Executive Order on Artificial Intelligence for Congress." Washington DC: Congressional Research Service, April 3, 2024. <https://crsreports.congress.gov/product/pdf/R/R47843#:~:text=On%20October%2030%2C%202023%2C%20the%20Biden%20Administration%20released,regulation%20of%20industry%2C%20and%20engagement%20with%20international%20partners>.
82. "History of the Cambridge Analytica Controversy | Bipartisan Policy Center." Accessed September 23, 2024. <https://bipartisanpolicy.org/blog/cambridge-analytica-controversy/>.
83. "How AI Plays a Role in Both Stopping and Committing DDoS Attacks." Accessed September 23, 2024. <https://securityintelligence.com/fight-fire-with-fire-how-ai-plays-a-role-in-both-stopping-and-committing-ddos-attacks/>.
84. "How Artificial Intelligence Influences Elections, and What We Can Do About It | Campaign Legal Center." Accessed September 23, 2024. <https://campaignlegal.org/update/how-artificial-intelligence-influences-elections-and-what-we-can-do-about-it>.
85. "How Election Experts Are Thinking about AI and Its Impact on the 2024 Elections | Hub." Accessed September 23, 2024. <https://hub.jhu.edu/2024/06/05/election-experts-artificial-intelligence/>.
86. H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020 (2020). <https://www.congress.gov/bill/116th-congress/house-bill/6216>.
87. Hung, Chen-Ling, Wen-Cheng Fu, Chang-Ce Liu, and Hui-Ju Tsai. "AI Disinformation Attacks and Taiwan's Responses during the 2024 Presidential Election." *Thomson Foundation*, n.d.
88. Huttenlocher, Dan, Asu Ozdaglar, and David Goldston. "A Framework for U.S. AI Governance: Creating a Safe and Thriving AI Sector," November 28, 2023.
89. Holloway, Andrew. "CISA Releases Findings from Its AI Pilot Program on Detecting Critical Vulnerabilities." *Alston & Bird*, September 26, 2024. [https://www.alstonprivacy.com/cisa-releases-findings-from-its-ai-pilot-program-on-detecting-critical-vulnerabilities/#:~:text=On%20July%2028%2C%202024%2C%20the,required%20by%20Executive%20Order%20\(EO\)](https://www.alstonprivacy.com/cisa-releases-findings-from-its-ai-pilot-program-on-detecting-critical-vulnerabilities/#:~:text=On%20July%2028%2C%202024%2C%20the,required%20by%20Executive%20Order%20(EO)).
90. IAPP Research and Insights. "Global AI Legislation Tracker," January 2024.

91. "Improving Model Safety Behavior with Rule-Based Rewards." Accessed September 23, 2024. <https://openai.com/index/improving-model-safety-behavior-with-rule-based-rewards/>.
92. "Imran Khan—Pakistan's Jailed Ex-PM—Uses Deepfake To Address Supporters." Accessed September 23, 2024. <https://www.forbes.com/sites/siladityaray/2023/12/18/imran-khan-pakistans-jailed-ex-leader-uses-ai-deepfake-to-address-online-election-rally/>.
93. "Inside Biden's Historic Decision to Drop out of the 2024 Race." Accessed September 23, 2024. <https://www.nbcnews.com/politics/joe-biden/bidens-historic-decision-drop-2024-race-rcna162930>.
94. "INTERNATIONAL STANDARD ISO/IEC 23894: Information Technology — Artificial Intelligence — Guidance on Risk Management," 2023. <https://cdn.standards.iteh.ai/samples/77304/cb803ee4e9624430a5db177459158b24/ISO-IEC-23894-2023.pdf>.
95. "ISO/IEC 42001:2023(En), Information Technology — Artificial Intelligence — Management System," 2023. <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en>.
96. John, Mark, and Sumanta Sen. "Elections in 2024 Are Going to Reshape Global Politics." *Reuters*, July 9, 2024. <https://www.reuters.com/graphics/GLOBAL-ELECTIONS2024/gdvzmkejkw/>.
97. Jones, Brian. "US Election Guide: Where Candidates Stand on Tech." *TechTarget*, October 31, 2024. [https://www.techtarget.com/searchcio/feature/US-election-guide-Where-candidates-stand-on-tech?utm\\_campaign=20241031\\_ERU-ACTIVE\\_WITHIN\\_90\\_DAYS&utm\\_medium=email&utm\\_source=SGERU&source\\_ad\\_id=366558634&src=15011495&asrc=EM\\_SGERU\\_304162092](https://www.techtarget.com/searchcio/feature/US-election-guide-Where-candidates-stand-on-tech?utm_campaign=20241031_ERU-ACTIVE_WITHIN_90_DAYS&utm_medium=email&utm_source=SGERU&source_ad_id=366558634&src=15011495&asrc=EM_SGERU_304162092).
98. Kerry. "NIST's AI Risk Management Framework Plants a Flag in the AI Debate." Brookings, February 15, 2023. <https://www.brookings.edu/articles/nists-ai-risk-management-framework-plants-a-flag-in-the-ai-debate/>.
99. Ladd, Valdez. "ISO/IEC 42001:2023: AI Governance and Operational Excellence." *Medium* (blog), March 25, 2024. [https://medium.com/@oracle\\_43885/iso-iec-42001-2023-ai-governance-and-operational-excellence-6b91e885f228](https://medium.com/@oracle_43885/iso-iec-42001-2023-ai-governance-and-operational-excellence-6b91e885f228).
100. Larsen, Benjamin. "The Geopolitics of AI and the Rise of Digital Sovereignty." Brookings, December 8, 2022. <https://www.brookings.edu/articles/the-geopolitics-of-ai-and-the-rise-of-digital-sovereignty/>.
101. Lee, Serena. "No Safety Without Standards: Defining Protocols for AI Red-Teaming Disclosures | TechPolicy.Press." Tech Policy Press, March 26, 2024. <https://techpolicy.press/no-safety-without-standards-defining-protocols-for-ai-redteaming-disclosures>.
102. Lerman, Rachel. "Trump's AI Executive Order: What to Know About Regulations and Military Use." *The Washington Post*, July 16, 2024. <https://www.washingtonpost.com/technology/2024/07/16/trump-ai-executive-order-regulations-military/>.
103. Lin, Belle. "AI Regulation Is Coming. Fortune 500 Companies Are Bracing for Impact." *Wall Street Journal*, August 27, 2024, sec. C Suite. <https://www.wsj.com/articles/ai-regulation-is-coming-fortune-500-companies-are-bracing-for-impact-94bba201>.
104. McCarthy, K. "AI Bill of Rights: What You Need to Know." Built In, October 4, 2023. <https://builtin.com/artificial-intelligence/ai-bill-of-rights#:~:text=Officially%20called%20the%20Blueprint%20for,companies%20like%20Microsoft%20and%20Google>.
105. McManus, Doyle. "California's New AI Regulations Are a Step in the Right Direction." *Los Angeles Times*, September 19, 2024. <https://www.latimes.com/opinion/story/2024-09-19/ai-artificial-intelligence-california-regulation>.
106. McCoy, Terrence. "Russia Targets Kamala Harris with Attack Ads Ahead of 2024 Election." *USA Today*, September 23, 2024. <https://www.usatoday.com/story/news/politics/elections/2024/09/23/russia-harris-attack-ads/75353125007/>.

107. Caroline Meinhardt, Christie M. Lawrence, Lindsey A. Gailmard, Daniel Zhang, Rishi Bommasani, Rohini Kosoglu, Peter Henderson, Russell Wald, and Daniel E. Ho. "By the Numbers: Tracking The AI Executive Order," November 16, 2023. <https://hai.stanford.edu/news/numbers-tracking-ai-executive-order>.
108. Miller, Gabby. "NIST Unveils Draft Guidance Reports Following Biden's AI Executive Order | TechPolicy.Press." Tech Policy Press, May 3, 2024. <https://techpolicy.press/nist-unveils-ai-draft-guidance-reports>.
109. MITRE. *MITRE Atlas*. Accessed September 27, 2024. <https://atlas.mitre.org/>.
110. MITRE. "MITRE and Microsoft Collaborate to Address Generative AI Security Risks." News Release, September 26, 2023. <https://www.mitre.org/news-insights/news-release/mitre-and-microsoft-collaborate-address-generative-ai-security-risks>.
111. MITRE. "MITRE Opens New AI Assurance and Discovery Lab." News Release, September 6, 2023. <https://www.mitre.org/news-insights/news-release/mitre-opens-new-ai-assurance-and-discovery-lab>.
112. Mola, Tessa. "Spamouflage: Chinese Network of Fake Social Media Accounts Disrupted." *Axios*, September 3, 2024. <https://www.axios.com/2024/09/03/spamouflage-chinese-network-fake-social-media-accounts>.
113. MongoDB. "The AI Stack: Understanding the Components of AI." Accessed October 29, 2024. <https://www.mongodb.com/resources/basics/artificial-intelligence/ai-stack>.
114. "Moldova Fights to Free Itself from Russia's AI-Powered Disinformation Machine – POLITICO." Accessed September 23, 2024. <https://www.politico.eu/article/moldova-fights-free-from-russia-ai-power-disinformation-machine-maia-sandu/>.
115. Mukherjee, Amitav. "Beyond ISO 42001: The Role of ISO/IEC 23894 in AI Risk Management." *Medium* (blog), June 11, 2024. <https://medium.com/@mukherjee.amitav/beyond-iso-42001-the-role-of-iso-iec-23894-in-ai-risk-management-7c4f3036544f>.
116. Multistate AI. "Artificial Intelligence (AI) Legislation," 2024. [https://www.multistate.ai/artificial-intelligence-ai-legislation?mkt\\_tok=MzgyLUUpaQi03OTgAAAGT2oF-eTclaoioMzq3IElwzLljj\\_102U\\_LRJ9PwfDTA2BV17W1awjvyEFhFTV1KknGLyHalvXF6AET-tfNzaMHRmThpxh9VgCX98v8knmyQAaA](https://www.multistate.ai/artificial-intelligence-ai-legislation?mkt_tok=MzgyLUUpaQi03OTgAAAGT2oF-eTclaoioMzq3IElwzLljj_102U_LRJ9PwfDTA2BV17W1awjvyEFhFTV1KknGLyHalvXF6AET-tfNzaMHRmThpxh9VgCX98v8knmyQAaA).
117. Multistate AI "AI Policy Overview: Colorado." Accessed August 31, 2024. <https://www.multistate.ai/ai-policy-overview-colorado>.
118. Multistate AI "AI Policy Overview: Florida." Accessed August 31, 2024. <https://www.multistate.ai/ai-policy-overview-florida>.
119. Multistate AI "AI Policy Overview: New York." Accessed August 31, 2024. <https://www.multistate.ai/ai-policy-overview-new-york>.
120. Muro, Mark, and Robert Maxim. "What Does the 2024 Election Mean for the Future of AI Governance?" Brookings, October 19, 2023. <https://www.brookings.edu/articles/what-does-the-2024-election-mean-for-the-future-of-ai-governance/>.
121. National Academies of Sciences, Engineering, and Medicine. *An Evidence Framework for Genetic Testing*. Washington, DC: The National Academies Press, 2017. <https://nap.nationalacademies.org/read/25120/chapter/7>.
122. National Conference of State Legislatures. "Artificial Intelligence 2024 Legislation," June 3, 2024. <https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation>.
123. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "A Plan for Global Engagement on AI Standards," April 2024. <https://airc.nist.gov/docs/NIST.AI.100-5.Global-Plan.ipd.pdf>.
124. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.
125. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," April 2024. <https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf>.

126. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "Crosswalk 1 – Terminology NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB)," n.d.
127. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency," April 2024. <https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf>.
128. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "Reva Schwartz." Accessed September 27, 2024. <https://www.nist.gov/people/reva-schwartz>.
129. National Institute of Standards and Technology U.S. Department of Commerce (NIST). "U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research." News Release, August 14, 2024. <https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research>.
130. Neenan, Alexandra G., and Kelley M. Saylor. "The AI Executive Order and Its Potential Implications for DOD." Washington DC: Congressional Research Service, December 12, 2023.
131. Neill, Bridget, John Hallmark, and Dan Diasio. "Key Takeaways from the Biden Administration Executive Order on AI," October 31, 2023. [https://www.ey.com/en\\_us/insights/public-policy/key-takeaways-from-the-biden-administration-executive-order-on-ai](https://www.ey.com/en_us/insights/public-policy/key-takeaways-from-the-biden-administration-executive-order-on-ai).
132. Ng, Alfred. "NSA report discloses Russian hacking days before US election." *CNET*. June 6, 2017. <https://www.cnet.com/news/privacy/nsa-russian-hacking-leaked-report-election-us/>.
133. Nichols, Anna Liz. "Michigan GOP Congressional Candidate Blames Fake MLK Endorsement Video on Campaign Volunteer • Michigan Advance." *Michigan Advance* (blog), June 15, 2024. <https://michiganadvance.com/2024/06/15/michigan-gop-congressional-candidate-blames-fake-mlk-endorsement-video-on-campaign-volunteer/>.
134. NIST AI Challenge Problems. "GenAI - Evaluating Generative AI," 2024. <https://ai-challenges.nist.gov/genai>.
135. "NIST Draft AI Guidance, Report, and Global Plan." Accessed July 30, 2024. <https://kpmg.com/us/en/articles/2024/nist-draft-ai-guidance-report-and-global-plan-reg-alert.html>.
136. Northwestern University. "The New Dawn of AI Evaluation: NIST's ARIA." CASMI, January 31, 2024. <https://casmi.northwestern.edu/news/articles/2024/the-new-dawn-of-ai-evaluation-nists-aria.html>.
137. NYC Consumer and Worker Protection. "Automated Employment Decision Tools: Frequently Asked Questions," June 29, 2023. <https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf>.
138. Office for Civil Rights (OCR) and US Department of Health and Human Services. "Section 1557 of the Patient Protection and Affordable Care Act." Text, July 22, 2010. <https://www.hhs.gov/civil-rights/for-individuals/section-1557/index.html>.
139. Office of Science and Technology Policy. "FACT SHEET: Biden-Harris Administration Announces Key Actions to Advance Tech Accountability and Protect the Rights of the American Public." The White House, October 4, 2022. <https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/>.
140. Office of the Director of National Intelligence and Admin. "INTEL - Artificial Intelligence Ethics Framework for the Intelligence Community." Accessed August 29, 2024. <https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community>.
141. Office of the Director of National Intelligence (ODNI), and Admin. "INTEL - Principles of Artificial Intelligence Ethics for the Intelligence Community." Accessed August 29, 2024. <https://www.intelligence.gov/principles-of-artificial-intelligence-ethics-for-the-intelligence-community>.
142. Office of the Director of National Intelligence. "Election Security Update as of Late July 2024." Last modified July 31, 2024. <https://www.odni.gov/index.php/fmic-news/3973-election-security-update-as-of-late-july-2024>.

143. Office of the Director of National Intelligence (ODNI). Election Security Update. September 23, 2024. <https://www.odni.gov/files/FMIC/documents/ODNI-Election-Security-Update-20240923.pdf>.
144. Office of Science and Technology Policy. *Artificial Intelligence and Quantum Information Science R&D Summary*. August 2020. <https://trumpwhitehouse.archives.gov/wp-content/uploads/2017/12/Artificial-Intelligence-Quantum-Information-Science-R-D-Summary-August-2020.pdf>.
145. Office of Science and Technology Policy. "Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government." Last modified December 3, 2020. <https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/>.
146. OpenAI. *Influence and Cyber Operations: An Update*. October 2024. [https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update\\_October-2024.pdf](https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf).
147. Pearcy, Sam. "HiddenLayer Research | A Guide to AI Red Teaming." HiddenLayer | Security for AI, June 20, 2024. <https://hiddenlayer.com/research/a-guide-to-ai-red-teaming/>.
148. Peatworks. "AI & Disability Inclusion Toolkit." Accessed July 30, 2024. <https://www.peatworks.org/ai-disability-inclusion-toolkit/>.
149. Peatworks. "The Equitable AI Playbook." Accessed July 30, 2024. <https://www.peatworks.org/ai-disability-inclusion-toolkit/the-equitable-ai-playbook/>.
150. Perkins Coie. "States Begin To Regulate AI in Absence of Federal Legislation," May 22, 2024. <https://www.perkinscoie.com/en/news-insights/states-begin-to-regulate-ai-in-absence-of-federal-legislation.html>.
151. Phillips. "Regulating AI in Healthcare: A Look at Biden's Executive Order." IMO Health, January 26, 2024. <https://www.imohealth.com/ideas/article/regulating-ai-in-healthcare-a-look-at-bidens-executive-order/>.
152. "Preparing to Fight AI-Backed Voter Suppression | Brennan Center for Justice." Accessed September 23, 2024. <https://www.brennancenter.org/our-work/research-reports/preparing-fight-ai-backed-voter-suppression>.
153. "Presidio AI Framework: Towards Safe Generative AI Models." AI Governance Alliance Briefing Paper Series 2024, 2024. [https://www3.weforum.org/docs/WEF\\_Presidio\\_AI%20Framework\\_2024.pdf](https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf).
154. R Street Institute. "Impact of Artificial Intelligence on Elections." Accessed September 23, 2024. <https://www.rstreet.org/research/impact-of-artificial-intelligence-on-elections/>.
155. Radware. "Iran's AI-Driven Social Media Botnets." Last modified August 21, 2024. <https://www.radware.com/security/threat-advisories-and-attack-reports/irans-ai-driven-social-media-botnets/>.
156. "Risk in Focus: Generative A.I. and the 2024 Election Cycle | CISA." Accessed September 23, 2024. <https://www.cisa.gov/resources-tools/resources/risk-focus-generative-ai-and-2024-election-cycle>.
157. Rizzo, Salvador. "OpenAI's Influence in Iran and ChatGPT's Role in the 2024 Election: Harris and Trump." *The Washington Post*, August 16, 2024. <https://www.washingtonpost.com/technology/2024/08/16/openai-influence-iran-chatgpt-election-harris-trump/>.
158. Roski, Joachim, Ezekiel J. Maier, Kevin Vigilante, Elizabeth A. Kane, and Michael E. Matheny. "Enhancing Trust in AI through Industry Self-Governance." *Journal of the American Medical Informatics Association* 28, no. 7 (April 2021): 1582–90.
159. Rumbaugh, Lauren. "Microsoft Skeleton Key Attacks Consistently Jailbreak AI Models, Allows Users to Directly Ask Forbidden Questions." *CPO Magazine*, September 21, 2024. <https://www.cpomagazine.com/cyber-security/microsoft-skeleton-key-attacks-consistently-jailbreak-ai-models-allows-users-to-directly-ask-forbidden-questions/>.
160. Seitz, Amanda and Barbara Ortutay. "Pennsylvania emerges as online misinformation hot spot." *AP News*, November 3, 2020. <https://apnews.com/article/pennsylvania-misinformation-hotspot-5d6a72bed293d3463ee6b83286ead1b6>.

161. Sólymos, Karin Kóváry. "Slovak Election Targeted by Pro-Kremlin Deepfake Hoax." *VSquare.Org* (blog), November 8, 2023. <https://vsquare.org/slovak-election-targeted-by-pro-kremlin-deepfake-hoax/>.
162. "S'pore Seeks International Feedback on New Governance Framework for Generative AI | The Straits Times." Accessed September 23, 2024. <https://www.straitstimes.com/singapore/s-pore-seeks-international-feedback-on-new-governance-framework-for-generative-ai>.
163. Stanford Institute for Human-Centered AI (HAI), Stanford RegLab, and Stanford Center for Research on Foundation Models (CRFM). "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER," June 18, 2024.
164. Stanford University Human-Centered Artificial Intelligence. "CHAPTER 7: Policy and Governance." In *Artificial Intelligence Index Report 2024*, 2024.
165. Sullivan, Kaylee. "Trump Vows to Cancel Biden Executive Order on AI to Protect Free Speech." *Washington Examiner*, July 26, 2024. <https://www.washingtonexaminer.com/news/2432277/trump-vows-to-cancel-biden-executive-order-on-ai-to-protect-free-speech/>.
166. Team, NIST AIRC. "NIST AIRC - Crosswalk Documents." Accessed July 30, 2024. [https://airc.nist.gov/AI\\_RMFM\\_Knowledge\\_Base/Crosswalks](https://airc.nist.gov/AI_RMFM_Knowledge_Base/Crosswalks).
167. "Technologists Wanted | Consumer Financial Protection Bureau." Accessed August 29, 2024. <https://www.consumerfinance.gov/about-us/blog/technologists-wanted/>.
168. "Test, Evaluation & Red-Teaming." *NIST*, December 21, 2023. <https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test>.
169. "The Good, the Bad and the Ugly Uses of Machine Learning in Election Campaigns | Centre For Public Impact (CPI)." Accessed September 23, 2024. <https://www.centreforpublicimpact.org/insights/good-bad-ugly-uses-machine-learning-election-campaigns>.
170. The White House. "Blueprint for an AI Bill of Rights," October 2022. <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>.
171. The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," October 30, 2023. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
172. The White House. "President Trump's FY 2021 Budget Commits to Double Investments in Key Industries." February 10, 2020. <https://trumpwhitehouse.archives.gov/briefings-statements/president-trumps-fy-2021-budget-commits-double-investments-key-industries-future/>.
173. Trivedi, Prem M. "Unpacking the White House's Executive Order on AI." *New America*, November 10, 2023. <http://newamerica.org/oti/blog/unpacking-the-white-houses-executive-order-on-ai/>.
174. United States Department of State. "Risk Management Profile for AI and Human Rights." *United States Department of State* (blog). Accessed July 30, 2024. <https://www.state.gov/risk-management-profile-for-ai-and-human-rights/>.
175. United Nations. "Can Artificial Intelligence (AI) Influence Elections?" *United Nations Western Europe*, June 7, 2024. <https://unric.org/en/can-artificial-intelligence-ai-influence-elections/>.
176. U.S. Agency for International Development. "Artificial Intelligence Action Plan," July 22, 2024. <https://www.usaid.gov/digital-development/artificial-intelligence-action-plan>.
177. "U.S. Artificial Intelligence Safety Institute." *NIST*, October 26, 2023. <https://www.nist.gov/aisi>.
178. U.S. Department of Commerce. "Department of Commerce Announces New Actions to Implement President Biden's Executive Order on AI," April 29, 2024. <https://www.commerce.gov/news/press-releases/2024/04/department-commerce-announces-new-actions-implement-president-bidens>.
179. U.S. Department of Defense. "DOD Adopts Ethical Principles for Artificial Intelligence." Accessed November 28, 2024. [DOD Adopts Ethical Principles for Artificial Intelligence > U.S. Department of Defense > Release](https://www.defense.gov/Newsroom/Releases/2024/11/28/dod-adopts-ethical-principles-for-artificial-intelligence/).
180. U.S Department of Defense. "Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022. <https://media.defense.gov/2024/Oct/26/2003571790/-1/-1/0/2024-06-RAI-STRATEGY-IMPLEMENTATION-PATHWAY.PDF>



181. US Department of Housing and Urban Development. "HUD Issues Fair Housing Act Guidance on Applications of Artificial Intelligence." HUD.gov / U.S. Department of Housing and Urban Development (HUD), May 2, 2024. [https://www.hud.gov/press/press\\_releases\\_media\\_advisories/hud\\_no\\_24\\_098](https://www.hud.gov/press/press_releases_media_advisories/hud_no_24_098).
182. US Department of Justice Civil Rights Division. "Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring." ADA.gov, July 26, 2024. <https://www.ada.gov/resources/ai-guidance/>.
183. U.S. Department of Labor. "HIRE Initiative." Accessed August 29, 2024. <https://www.dol.gov/agencies/ofccp/Hire-Initiative>.
184. US Department of Treasury. "Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector," March 2024. <https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf>.
185. U.S. Election Assistance Commission. "Alert: Misleading Voter Registration Phishing Email." Last modified October 11, 2024. Accessed October 29, 2024. <https://www.eac.gov/news/2024/10/11/alert-misleading-voter-registration-phishing-email>.
186. US Equal Employment Opportunity Commission. "U.S. EEOC and U.S. Department of Justice Warn against Disability Discrimination," May 12, 2022. <https://www.eeoc.gov/newsroom/us-eeoc-and-us-department-justice-warn-against-disability-discrimination>.
187. Vasquez, Christian. "Three Bills Governing AI in Elections Pass Senate Committee." *CyberScoop* (blog), May 15, 2024. <https://cyberscoop.com/klobuchar-ai-election-bill-senate-markup/>.
188. Weil, Gabriel. "The Pros and Cons of California's Proposed SB-1047 AI Safety Law." Default, May 8, 2024. <https://www.lawfaremedia.org/article/california-s-proposed-sb-1047-would-be-a-major-step-forward-for-ai-safety-but-there-s-still-room-for-improvement>.
189. "What Role Is AI Playing in Election Disinformation? | Brookings." Accessed July 11, 2024. <https://www.brookings.edu/articles/what-role-is-ai-playing-in-election-disinformation/>.
190. Wheeler, T. N. "Trump Pledges to Ax Biden's AI Executive Order." *Nextgov*, July 26, 2024. <https://www.nextgov.com/artificial-intelligence/2024/07/trump-pledges-ax-bidens-ai-executive-order/397905/>.
191. White & Case LLP. "AI Watch: Global Regulatory Tracker - United States," May 13, 2024. <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states>.
192. Wood, Derek. "9 Principles of an AI Governance Framework." Duality Technologies, April 4, 2024. <https://dualitytech.com/blog/ai-governance-framework/>.
193. World Economic Forum. "AI Governance Alliance Briefing Paper Series," January 2024. [https://www3.weforum.org/docs/WEF\\_AI\\_Governance\\_Alliance\\_Briefing\\_Paper\\_Series\\_2024.pdf](https://www3.weforum.org/docs/WEF_AI_Governance_Alliance_Briefing_Paper_Series_2024.pdf).
194. World Economic Forum. *Generative AI Governance: A Framework for the Future*. Geneva: World Economic Forum, 2024. [https://www3.weforum.org/docs/WEF\\_Generative\\_AI\\_Governance\\_2024.pdf](https://www3.weforum.org/docs/WEF_Generative_AI_Governance_2024.pdf). World Economic Forum. "What's in the US 'AI Bill of Rights' - and What Isn't," October 14, 2022. <https://www.weforum.org/agenda/2022/10/understanding-the-ai-bill-of-rights-protection/>.

## 13. Footnotes

- <sup>1</sup> McKinsey, 'The Economic Potential of Generative AI: The next Productivity Frontier'.
- <sup>2</sup> Aspen Digital. "Generative AI Regulation and Cybersecurity." Aspen Institute, 2024.
- <sup>3</sup> John, Mark, and Sumanta Sen. "Elections in 2024 Are Going to Reshape Global Politics."
- <sup>4</sup> Ibid.
- <sup>5</sup> Lin, Belle. "AI Regulation Is Coming. Fortune 500 Companies Are Bracing for Impact."
- <sup>6</sup> Ash Center. "AI and the 2024 Elections," May 29, 2024.
- <sup>7</sup> NIST, "U.S. Artificial Intelligence Safety Institute," October 26, 2023.
- <sup>8</sup> Hoffman and Kim, 'Reducing the Risks of Artificial Intelligence for Military Decision Advantage'.
- <sup>9</sup> CBS News, "Generative AI Poses Threat to Election Security, Federal Intelligence Agencies Warn."
- <sup>10</sup> AP News, "Clinton Wins Popular Vote by Nearly 2.9 Million."
- <sup>11</sup> Ball, Molly. "How COVID-19 Changed Everything About the 2020 Election," August 6, 2020.
- <sup>12</sup> NBC News, "Inside Biden's Historic Decision to Drop out of the 2024 Race."
- <sup>13</sup> Ibid.
- <sup>14</sup> "Did Russia Influence Brexit? | Brexit Bits, Bobs, and Blogs." *CSIS*.
- <sup>15</sup> Bipartisan Policy Center. "History of the Cambridge Analytica Controversy."
- <sup>16</sup> 120. Seitz, Amanda and Barbara Ortutay. "Pennsylvania emerges as online misinformation hot spot." AP News, November 3, 2020.
- <sup>17</sup> NBC News, "Conspiracy Theories That Biden Is Dead, Dying Boosted by Musk's X."
- <sup>18</sup> The Diplomat, "AI and India's General Elections."
- <sup>19</sup> Egan, Matt, "Exclusive: 42% of CEOs Say AI Could Destroy Humanity in Five to Ten Years."
- <sup>20</sup> Hung, Chen-Ling, Wen-Cheng Fu, Chang-Ce Liu, and Hui-Ju Tsai, "AI Disinformation Attacks and Taiwan's Responses during the 2024 Presidential Election." *Thomson Foundation*.
- <sup>21</sup> Brennan Center, "Preparing to Fight AI-Backed Voter Suppression."
- <sup>22</sup> Cybersecurity & Infrastructure Security Agency (CISA). "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024.
- <sup>23</sup> Ibid.
- <sup>24</sup> Ibid.
- <sup>25</sup> National Coordinator for Critical Infrastructure Security and Resilience. "Tactics of Disinformation."
- <sup>26</sup> Graphika, "Deepfake It Till You Make It."
- <sup>27</sup> Solyomos, Karin Kóváry. "Slovak Election Targeted by Pro-Kremlin Deepfake Hoax."
- <sup>28</sup> Brennan Center, "Deepfakes, Elections, and Shrinking the Liar's Dividend."
- <sup>29</sup> National Coordinator for Critical Infrastructure Security and Resilience. "Tactics of Disinformation."
- <sup>30</sup> NBC News, "Conspiracy Theories That Biden Is Dead, Dying Boosted by Musk's X."
- <sup>31</sup> CISA. "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024.
- <sup>32</sup> CISOMAG. "Artificial Intelligence as Security Solution and Weaponization by Hackers." *CISO MAG | Cyber Security Magazine (blog)*, December 9, 2019.
- <sup>33</sup> CISA. "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024.
- <sup>34</sup> Ibid.
- <sup>35</sup> National Coordinator for Critical Infrastructure Security and Resilience. "Tactics of Disinformation."
- <sup>36</sup> Brennan Center, "Preparing to Fight AI-Backed Voter Suppression." Brennan Center.
- <sup>37</sup> National Coordinator for Critical Infrastructure Security and Resilience. "Tactics of Disinformation."
- <sup>38</sup> Ibid.
- <sup>39</sup> CISA. "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024.
- <sup>40</sup> Brennan Center, "Preparing to Fight AI-Backed Voter Suppression."
- <sup>41</sup> Brennan Center, "Preparing to Fight AI-Backed Voter Suppression."
- <sup>42</sup> Harvard Business Review, "AI Will Increase the Quantity — and Quality — of Phishing Scams."
- <sup>43</sup> Ibid.
- <sup>44</sup> Ibid.
- <sup>45</sup> Security Intelligence, "How AI Plays a Role in Both Stopping and Committing DDoS Attacks."
- <sup>46</sup> Federal Bureau of Investigation, "DDoS Attacks: Could Hinder Access to Election Information, Would Not Prevent Voting."
- <sup>47</sup> Sharma, Shweta, "ChatGPT Creates Mutating Malware That Evades Detection by EDR."
- <sup>48</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."
- <sup>49</sup> Aspen Digital. "Envisioning Cyber Futures with A.I." Aspen Institute, 2024.
- <sup>50</sup> World Economic Forum, 'Global Risks Report 2024', 19; Woollacot, 'China Targets US Voters With New AI Misinformation Techniques'.
- <sup>51</sup> CISA, "Risk in Focus: Generative A.I. and the 2024 Election Cycle."
- <sup>52</sup> Ibid.
- <sup>53</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."
- <sup>54</sup> Hacker, Brittany, Roseberry, Alexandra, "AI Threats in Elections: What Nonprofits Must Know."
- <sup>55</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."
- <sup>56</sup> McIsaac, Chris, "Impact of Artificial Intelligence on Elections."
- <sup>57</sup> U.S. Election Assistance Commission, "Alert: Misleading Voter Registration Phishing Email."
- <sup>58</sup> "A Framework for Election Vendor Oversight | Brennan Center for Justice."
- <sup>59</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."
- <sup>60</sup> Ng, Alfred, "NSA report discloses Russian hacking days before US election," CNET, June 6, 2017.
- <sup>61</sup> National Academies of Sciences, Engineering, and Medicine, *An Evidence Framework for Genetic Testing*, Washington, DC: The National Academies Press, 2017.
- <sup>62</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."

- 
- <sup>63</sup> FBI, "Election Crimes."
- <sup>64</sup> Center for Internet Security. "Election Infrastructure Information Sharing and Analysis Center (EI-ISAC)."
- <sup>65</sup> CISA, "Government Coordinating Councils," Accessed September 26, 2024.
- <sup>66</sup> CISA, "Sector Coordinating Councils," Accessed September 26, 2024.
- <sup>67</sup> CISA, Physical Security Checklist for Election Offices, September 2024.
- <sup>68</sup> CISA, Readiness and Resilience Checklist for Election Offices, September 2024..
- <sup>69</sup> Ibid.
- <sup>70</sup> CISA, "Election Security," September 27, 2024.
- <sup>71</sup> CISA, "Risk in Focus: Generative A.A. and the 2024 Election Cycle."
- <sup>72</sup> U.S. Election Assistance Commission, "Alert: Misleading Voter Registration Phishing Email."
- <sup>73</sup> Ibid.
- <sup>74</sup> Ash Center. "AI and the 2024 Elections," May 29, 2024.
- <sup>75</sup> OpenAI, "Improving Model Safety Behavior with Rule-Based Rewards."
- <sup>76</sup> Ibid.
- <sup>77</sup> Smith, Travis, "A Guide to Redteaming."
- <sup>78</sup> Georgetown Center for Security and Emerging Technology, "What Does AI Red-Teaming Actually Mean?"
- <sup>79</sup> National Institute of Standards and Technology, "Test, Evaluation & Red-Teaming."
- <sup>80</sup> Lee, Serena. "No Safety Without Standards: Defining Protocols for AI Red-Teaming Disclosures | TechPolicy.Press."
- <sup>81</sup> NIST, "Reva Schwartz," Accessed September 27, 2024..
- <sup>82</sup> Northwestern University, "The New Dawn of AI Evaluation: NIST's ARIA," CASMI, January 31, 2024.
- <sup>83</sup> Johns Hopkins University, "How Election Experts Are Thinking about AI and Its Impact on the 2024 Elections."
- <sup>84</sup> Brennan Center, "Safeguards for Using Artificial Intelligence in Election Administration."
- <sup>85</sup> Centre For Public Impact (CPI). "The Good, the Bad and the Ugly Uses of Machine Learning in Election Campaigns."
- <sup>86</sup> United Nations Western Europe, "Can Artificial Intelligence (AI) Influence Elections?"
- <sup>87</sup> Centre For Public Impact (CPI). "The Good, the Bad and the Ugly Uses of Machine Learning in Election Campaigns."
- <sup>88</sup> Brennan Center, "Safeguards for Using Artificial Intelligence in Election Administration."
- <sup>89</sup> United Nations Western Europe, "Can Artificial Intelligence (AI) Influence Elections?"
- <sup>90</sup> Brennan Center, "Safeguards for Using Artificial Intelligence in Election Administration."
- <sup>91</sup> Ibid.
- <sup>92</sup> Ibid.
- <sup>93</sup> Centre For Public Impact (CPI). "The Good, the Bad and the Ugly Uses of Machine Learning in Election Campaigns."
- <sup>94</sup> Office of the Director of National Intelligence (ODNI), Election Security Update, September 23, 2024.
- <sup>95</sup> Dilanian, Ken, "Russia, Iran Using AI to Influence U.S. Election, DNI Warns," ABC News, September 25, 2024.
- <sup>96</sup> Ash Center, "AI and the 2024 Elections."
- <sup>97</sup> Brookings. "What Role Is AI Playing in Election Disinformation?"
- <sup>98</sup> "Deepfakes, Elections, and Shrinking the Liar's Dividend | Brennan Center for Justice," February 8, 2024.
- <sup>99</sup> Aspen Digital, "Generative AI Regulation and Cybersecurity," 5. Interview Findings, 8-18 July.
- <sup>100</sup> World Economic Forum, "AI Governance Alliance Briefing Paper Series," 44.
- <sup>101</sup> World Economic Forum, 45.
- <sup>102</sup> World Economic Forum, 45.
- <sup>103</sup> World Economic Forum, 45.
- <sup>104</sup> World Economic Forum, *Generative AI Governance: A Framework for the Future*, Geneva: World Economic Forum, 2024. 6.
- <sup>105</sup> Stanford University Human-Centered Artificial Intelligence, "CHAPTER 7: Policy and Governance."
- <sup>106</sup> Vasquez, "Three Bills Governing AI in Elections Pass Senate Committee."
- <sup>107</sup> Brennan Center for Justice, "Artificial Intelligence Legislation Tracker."
- <sup>108</sup> DLA Piper, "AI Legislation Advances in U.S. House of Representatives," Last modified October 2024, Accessed October 29, 2024.
- <sup>109</sup> The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."
- <sup>110</sup> Neill, Hallmark, and Diasio, "Key Takeaways from the Biden Administration Executive Order on AI."
- <sup>111</sup> Meinhardt et al., "By the Numbers."
- <sup>112</sup> Meinhardt et al.
- <sup>113</sup> Stanford Institute for Human-Centered AI (HAI), Stanford RegLab, and Stanford Center for Research on Foundation Models (CRFM), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>114</sup> According to the Executive Office of the President, current as of June 2024.
- <sup>115</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>116</sup> According to the Department of Commerce as of June 2024.
- <sup>117</sup> "GenAI - Evaluating Generative AI."
- <sup>118</sup> "GenAI - Evaluating Generative AI."
- <sup>119</sup> DOC, "Department of Commerce Announces New Actions to Implement President Biden's Executive Order on AI."
- <sup>120</sup> U.S. Department of Commerce.
- <sup>121</sup> "Department of Commerce Proposes Rule on IaaS Product-Related Customer Identification and AI-Related Reporting Requirements | Advisories."
- <sup>122</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>123</sup> According to DHS, current as of June 2024.
- <sup>124</sup> DHS, "DHS Publishes Guidelines and Report to Secure Critical Infrastructure and Weapons of Mass Destruction from AI-Related Threats."
- <sup>125</sup> DHS, "Promoting AI Safety and Security."
- <sup>126</sup> DHS, "Fact Sheet: DHS Facilitates Safe and Responsible Deployment and Use of Artificial Intelligence," last modified April 29, 2024, accessed October 29, 2024.
- <sup>127</sup> Holloway, Andrew, "CISA Releases Findings from Its AI Pilot Program on Detecting Critical Vulnerabilities," Alston & Bird, September 26, 2024.
- <sup>128</sup> CISA, "CISA Roadmap for Artificial Intelligence."

- 
- <sup>129</sup> CISA, 3.
- <sup>130</sup> CISA, "CISA, JCDC, Government and Industry Partners Conduct AI Tabletop Exercise."
- <sup>131</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>132</sup> Based on Status According to the OPM, current as of June 2024.
- <sup>133</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>134</sup> Based on Status According to Department of State, current as of June 2024.
- <sup>135</sup> United States Department of State, "Risk Management Profile for AI and Human Rights."
- <sup>136</sup> United States Department of State.
- <sup>137</sup> United States Department of State.
- <sup>138</sup> Phillips, "Regulating AI in Healthcare."
- <sup>139</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>140</sup> Based on Status According to Department of State, current as of June 2024.
- <sup>141</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>142</sup> Based on Status According to the DOE, current as of June 2024.
- <sup>143</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>144</sup> Based on Status According to the NSF, current as of June 2024.
- <sup>145</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>146</sup> Based on Status According to the DOJ, current as of June 2024.
- <sup>147</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>148</sup> Based on Status According to the DOL, current as of June 2024.
- <sup>149</sup> Stanford Institute for Human-Centered AI (HAI), "THE SAFE, SECURE, AND TRUSTWORTHY AI EO TRACKER."
- <sup>150</sup> Stanford Institute for Human-Centered AI (HAI), Stanford RegLab, and Stanford Center for Research on Foundation Models (CRFM).
- <sup>151</sup> Neenan and Saylor, "The AI Executive Order and Its Potential Implications for DOD."
- <sup>152</sup> Neenan and Saylor.
- <sup>153</sup> The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."
- <sup>154</sup> US Department of Treasury, "Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector."
- <sup>155</sup> Harris, "Highlights of the 2023 Executive Order on Artificial Intelligence for Congress."
- <sup>156</sup> Trivedi, "Unpacking the White House's Executive Order on AI.," Interview Findings 8-18 July 2024.
- <sup>157</sup> Ibid.
- <sup>158</sup> Wheeler, T. N, "Trump Pledges to Ax Biden's AI Executive Order," Nextgov, July 26, 2024.
- <sup>159</sup> Sullivan, Kaylee, "Trump Vows to Cancel Biden Executive Order on AI to Protect Free Speech," Washington Examiner, July 26, 2024.
- <sup>160</sup> Jones, Brian, "US Election Guide: Where Candidates Stand on Tech," *TechTarget*, October 31, 2024.
- <sup>161</sup> Cohen, Richard, "Artificial Intelligence Regulations: Biden and Trump Need to Act," Chicago Tribune, June 10, 2024.
- <sup>162</sup> Lerman, Rachel, "Trump's AI Executive Order: What to Know About Regulations and Military Use," *The Washington Post*, July 16, 2024.
- <sup>163</sup> EOP, "Maintaining American Leadership in Artificial Intelligence," Federal Register 84, no. 30 (February 14, 2019).
- <sup>164</sup> Lerman, Rachel, "Trump's AI Executive Order: What to Know About Regulations and Military Use."
- <sup>165</sup> The White House, "President Trump's FY 2021 Budget Commits to Double Investments in Key Industries," February 10, 2020.
- <sup>166</sup> Office of Science and Technology Policy, Artificial Intelligence and Quantum Information Science R&D Summary, August 2020.
- <sup>167</sup> Office of Science and Technology Policy, "Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," Last modified December 3, 2020.
- <sup>168</sup> Muro, Mark, and Robert Maxim, "What Does the 2024 Election Mean for the Future of AI Governance?" Brookings, October 19, 2023.
- <sup>169</sup> Cohen, Richard, "Artificial Intelligence Regulations: Biden and Trump Need to Act."
- <sup>170</sup> Lerman, Rachel, "Trump's AI Executive Order: What to Know About Regulations and Military Use."
- <sup>171</sup> The White House, "Blueprint for an AI Bill of Rights."
- <sup>172</sup> Firth-Butterfield, Silverman, and Larsen, "Understanding the US 'AI Bill of Rights' - and How It Can Help Keep AI Accountable."
- <sup>173</sup> Firth-Butterfield, Silverman, and Larsen.
- <sup>174</sup> Firth-Butterfield, Silverman, and Larsen.
- <sup>175</sup> The White House, "Blueprint for an AI Bill of Rights."
- <sup>176</sup> Ibid.
- <sup>177</sup> Ibid.
- <sup>178</sup> Office of Science and Technology Policy, "FACT SHEET."
- <sup>179</sup> Goldman, "What the Blueprint for an AI Bill of Rights Means for Workers"; Freund, "How We're Ramping Up Our Enforcement of Surveillance Reporting."
- <sup>180</sup> US Department of Justice Civil Rights Division, "Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring"; "U.S. EEOC and U.S. Department of Justice Warn against Disability Discrimination."
- <sup>181</sup> "The Equitable AI Playbook"; "AI & Disability Inclusion Toolkit."
- <sup>182</sup> U.S. Department of Labor, "HIRE Initiative."
- <sup>183</sup> "CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms."
- <sup>184</sup> CDO Magazine Bureau, "US Department of Education Releases Guidelines for Integrating AI into Edtech."
- <sup>185</sup> Office for Civil Rights & US Department of Health and Human Services, "Section 1557 of the Patient Protection & Affordable Care Act."
- <sup>186</sup> US Department of Housing and Urban Development, "HUD Issues Fair Housing Act Guidance on Applications of Artificial Intelligence."
- <sup>187</sup> "Artificial Intelligence Action Plan."
- <sup>188</sup> "Federal AI Use Case Inventories."
- <sup>189</sup> "DOE AI Risk Management Playbook (AIRMP)."
- <sup>190</sup> "DOD Adopts Ethical Principles for Artificial Intelligence"; U.S DOD, "Responsible Artificial Intelligence Strategy and Implementation Pathway."
- <sup>191</sup> Office of the Director of National Intelligence and Admin, "INTEL - Principles of Artificial Intelligence Ethics for the Intelligence Community"; Intelligence and Admin, "INTEL - Artificial Intelligence Ethics Framework for the Intelligence Community."
- <sup>192</sup> Chia, "AI Bill of Rights"; Engler, "The AI Bill of Rights Makes Uneven Progress on Algorithmic Protections."
- <sup>193</sup> Engler, "The AI Bill of Rights Makes Uneven Progress on Algorithmic Protections."

- 
- <sup>194</sup> National Institute of Standards and Technology U.S. Department of Commerce (NIST), “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.”
- <sup>195</sup> National Institute of Standards and Technology U.S. Department of Commerce (NIST).
- <sup>196</sup> National Institute of Standards and Technology U.S. Department of Commerce (NIST).
- <sup>197</sup> National Institute of Standards and Technology U.S. Department of Commerce (NIST), “Crosswalk 1 – Terminology NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB).”
- <sup>198</sup> Various Interview Findings, 8-18 July, 2024; Kerry, “NIST’s AI Risk Management Framework Plants a Flag in the AI Debate.”
- <sup>199</sup> Various Interview Findings, 8-18 July, 2024; Kerry.
- <sup>200</sup> The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.”
- <sup>201</sup> Booth et al., “Secure Software Development Practices for Generative AI and Dual-Use Foundation Models.”
- <sup>202</sup> Booth et al., 4.
- <sup>203</sup> Booth et al., 4.
- <sup>204</sup> National Institute of Standards and Technology U.S. Department of Commerce (NIST), “A Plan for Global Engagement on AI Standards.”
- <sup>205</sup> NIST, “Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency.”
- <sup>206</sup> “ISO/IEC 42001:2023(En), Information Technology — Artificial Intelligence — Management System.”
- <sup>207</sup> Ladd, “ISO/IEC 42001.”
- <sup>208</sup> Ladd.
- <sup>209</sup> “ISO/IEC 42001:2023(En), Information Technology — Artificial Intelligence — Management System.”
- <sup>210</sup> Mukherjee, “Beyond ISO 42001.”
- <sup>211</sup> Mukherjee, “Beyond ISO 42001.”
- <sup>212</sup> Mukherjee.
- <sup>213</sup> Mukherjee, “Beyond ISO 42001.”
- <sup>214</sup> Mukherjee.
- <sup>215</sup> Multistate, “Artificial Intelligence (AI) Legislation”; National Conference of State Legislatures, “Artificial Intelligence 2024 Legislation.”
- <sup>216</sup> Bryan Cave Leighton Paisner LLP, “US State-by-State AI Legislation Snapshot.”
- <sup>217</sup> Multistate, “Artificial Intelligence (AI) Legislation”; National Conference of State Legislatures, “Artificial Intelligence 2024 Legislation.”
- <sup>218</sup> McManus, Doyle, “California’s New AI Regulations Are a Step in the Right Direction,” *Los Angeles Times*, September 19, 2024
- <sup>219</sup> World Economic Forum, “AI Governance Alliance Briefing Paper Series.”
- <sup>220</sup> World Economic Forum, 45.
- <sup>221</sup> World Economic Forum, 45.
- <sup>222</sup> Larsen, “The Geopolitics of AI and the Rise of Digital Sovereignty.”
- <sup>223</sup> Roski et al, 1854.
- <sup>224</sup> Roski et al.
- <sup>225</sup> MongoDB, “The AI Stack: Understanding the Components of AI.”
- <sup>226</sup> Huttenlocher, Ozdaglar, and Goldston, “A Framework for U.S. AI Governance: Creating a Safe and Thriving AI Sector,” 4.
- <sup>227</sup> Huttenlocher, Ozdaglar, and Goldston, 4.
- <sup>228</sup> Huttenlocher, Ozdaglar, and Goldston, 4.
- <sup>229</sup> Huttenlocher, Ozdaglar, and Goldston, 5.
- <sup>230</sup> Huttenlocher, Ozdaglar, and Goldston, 5.
- <sup>231</sup> Huttenlocher, Ozdaglar, and Goldston, 6.
- <sup>232</sup> MITRE, “MITRE Opens New AI Assurance and Discovery Lab,” News Release, September 6, 2023.
- <sup>233</sup> MITRE, “MITRE and Microsoft Collaborate to Address Generative AI Security Risks,” News Release, September 26, 2023.
- <sup>234</sup> MITRE. *MITRE Atlas*. Accessed September 27, 2024. <https://atlas.mitre.org/>.
- <sup>235</sup> Huttenlocher, Ozdaglar, and Goldston, 7.
- <sup>236</sup> Huttenlocher, Ozdaglar, and Goldston, 7.
- <sup>237</sup> Huttenlocher, Ozdaglar, and Goldston, 7.
- <sup>238</sup> The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.”
- <sup>239</sup> Huttenlocher, Ozdaglar, and Goldston, “A Framework for U.S. AI Governance: Creating a Safe and Thriving AI Sector.”
- <sup>240</sup> Huttenlocher, Ozdaglar, and Goldston, 1–4.
- <sup>241</sup> “Presidio AI Framework: Towards Safe Generative AI Models.”
- <sup>242</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 3.
- <sup>243</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 6–7.
- <sup>244</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 6.
- <sup>245</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 7.
- <sup>246</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 8.
- <sup>247</sup> *Ibid.*
- <sup>248</sup> *Ibid.*
- <sup>249</sup> *Ibid.*
- <sup>250</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 9.
- <sup>251</sup> *Ibid.*
- <sup>252</sup> *Ibid.*
- <sup>253</sup> *Ibid.*
- <sup>254</sup> *Ibid.*
- <sup>255</sup> *Ibid.*
- <sup>256</sup> *Ibid.*
- <sup>257</sup> *Ibid.*
- <sup>258</sup> *Ibid.*
- <sup>259</sup> “Presidio AI Framework: Towards Safe Generative AI Models,” 11.
- <sup>260</sup> *Ibid.*
- <sup>261</sup> Daniels and Murdick, “Enabling Principles for AI Governance.”
- <sup>262</sup> Daniels and Murdick, 4.

---

<sup>263</sup> Ibid.

<sup>264</sup> Ibid.

<sup>265</sup> Daniels and Murdick, 4–5.

<sup>266</sup> Ibid.

<sup>267</sup> Ibid.

<sup>268</sup> Daniels and Murdick, 5.

<sup>269</sup> Ibid.

<sup>270</sup> Daniels and Murdick, 6.

<sup>271</sup> Ibid.

<sup>272</sup> Ibid.

<sup>273</sup> Ibid.

<sup>274</sup> Ibid.

<sup>275</sup> Ibid.

<sup>276</sup> Ibid.

<sup>277</sup> Ibid.

<sup>278</sup> Daniels and Murdick, 8.

<sup>279</sup> Ibid.

<sup>280</sup> McCarthy, K, "AI Bill of Rights: What You Need to Know," Built In, October 4, 2023.

<sup>281</sup> NIST, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models. April 2024.

<sup>282</sup> NIST, "U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research," News Release, August 14, 2024.

<sup>283</sup> CISA. "Securing Election Infrastructure Against the Tactics of Foreign Malign Influence Operations," 2024.

<sup>284</sup> "AI and India's General Elections." *The Diplomat*.

<sup>285</sup> Ibid.

<sup>286</sup> Trajano, Karryl Kim; Jalli, Sagun Nuurrianti, "AI and Elections: Lessons for Southeast Asia."

<sup>287</sup> Han, Goh Yan, "'S'pore Seeks International Feedback on New Governance Framework for Generative AI."

<sup>288</sup> Trajano, Karryl Kim; Jalli, Sagun Nuurrianti, "AI and Elections: Lessons for Southeast Asia."

<sup>289</sup> NBC News, "Conspiracy Theories That Biden Is Dead, Dying Boosted by Musk's X."

<sup>290</sup> Dig Watch, "'Elections in 2024 and Global Politics."

<sup>291</sup> Associated Press, "'Biden Deepfake Spreads Online after Withdrawal from 2024 Race ."

<sup>292</sup> Nichols, Anna Liz. "Michigan GOP Congressional Candidate Blames Fake MLK Endorsement Video on Campaign Volunteer."

<sup>293</sup> Garver, Rob. "AI Chatbots Provide False Information About November Elections."

<sup>294</sup> Angwin, Julia, Alondra Nelson, and Rina Palta. "Seeking Reliable Election Information? Don't Trust AI"

<sup>295</sup> McCoy, Terrence, "Russia Targets Kamala Harris with Attack Ads Ahead of 2024 Election," USA Today, September 23, 2024.

<sup>296</sup> Council on Foreign Relations, "Deepfake of Kamala Harris Reups Questions on Tech's Self-Regulation."

<sup>297</sup> Rizzo, Salvador, "OpenAI's Influence in Iran & ChatGPT's Role in the 2024 Election: Harris and Trump," *The Washington Post*. Aug 2024.

<sup>298</sup> ODNI, "Election Security Update as of Late July 2024," Last modified July 31, 2024.

<sup>299</sup> Radware, "Iran's AI-Driven Social Media Botnets," Last modified August 21, 2024.

<sup>300</sup> Campbell, James. "Influence Actors Likely to Adjust Tactics Amid Election Chaos." *The Record*, September 24, 2024.

<sup>301</sup> OpenAI, *Influence and Cyber Operations: An Update*, October 2024.

<sup>302</sup> Rumbaugh, Lauren, "Microsoft Skeleton Key Attacks Consistently Jailbreak AI Models, Allows Users to Directly Ask Forbidden Questions," *CPO Magazine*, September 21, 2024.

<sup>303</sup> "Hackers Use AI-Generated Code to Develop Malware, Says HP Threat Report," *The Indian Express*, September 26, 2024.

<sup>304</sup> DFRLab, "FIMI 101: Understanding Foreign Influence Operations in the Information Environment," September 26, 2024.

<sup>305</sup> Campbell, James, "Spamouflage: The Influence Operation Behind China's Information Manipulation," *The Record*, September 25, 2024

<sup>306</sup> Mola, Tessa, "Spamouflage: Chinese Network of Fake Social Media Accounts Disrupted," *Axios*, September 3, 2024.

<sup>307</sup> Sóllymos, Karin Kőváry. "Slovak Election Targeted by Pro-Kremlin Deepfake Hoax."

<sup>308</sup> Franklin, Margarita, "Adversarial Threat Report."

<sup>309</sup> Scott, Mark, "Moldova fights to free itself from Russia's AI-powered disinformation machine."

<sup>310</sup> Ray, Saladitya, "Imran Khan—Pakistan's Jailed Ex-Leader—Uses AI Deepfake To Address Online Election Rally."

<sup>311</sup> OpenAI, *Influence and Cyber Operations: An Update*, October 2024.

<sup>312</sup> Aspen Digital, "Generative A.I. Regulation and Cybersecurity: A Global View of Policymaking"; "AI Risk Management Framework."

<sup>313</sup> Hardy and Marks, "California Senator Scott Wiener Introduced New Bill That Would Require a Kill Switch for Applicable Artificial Intelligence Models."

<sup>314</sup> McManus, Doyle, "California's New AI Regulations Are a Step in the Right Direction," *Los Angeles Times*, September 19, 2024

<sup>315</sup> Carnegie Endowment for International Peace, "California SB 1047: AI Safety Bill Veto Lessons," Last modified October 2024.

<sup>316</sup> Civil Code - CIV DIVISION 3. OBLIGATIONS [1427 - 3273.69] ( Heading of Division 3 amended by Stats. 1988, Ch. 160, Sec. 14. ) PART 4. OBLIGATIONS ARISING FROM PARTICULAR TRANSACTIONS [1738 - 3273.69] ( Part 4 enacted 1872. ); White & Case LLP, "AI Watch."

<sup>317</sup> "AI Policy Overview."

<sup>318</sup> "AI Policy Overview."

<sup>319</sup> "AI Policy Overview."

<sup>320</sup> "AI Policy Overview: Michigan."

<sup>321</sup> "AI Policy Overview."

<sup>322</sup> "AI Policy Overview."

<sup>323</sup> NYC Consumer and Worker Protection, "Automated Employment Decision Tools: Frequently Asked Questions."

<sup>324</sup> Law was enacted in 2021, took effect on January 1, 2023, and enforcement begun July 5, 2023.

<sup>325</sup> "AI Policy Overview."

<sup>326</sup> "AI Policy Overview."

<sup>327</sup> Weil, "The Pros and Cons of California's Proposed SB-1047 AI Safety Law."