Do AI Companies Make Good on Voluntary Commitments to the White House?

Jennifer Wang[1], **Kayla Huang**[2], **Kevin Klyman**[3], **Rishi Bommasani**[3]
[1]Brown University, Providence, RI, USA
[2]Harvard University, Cambridge, MA, USA
[3]Stanford University, Stanford, CA, USA
jennifer_wang2@brown.edu, kaylahuang@college.harvard.edu, kklyman@stanford.edu, nlprishi@stanford.edu

Abstract

Voluntary commitments are central to international AI governance, as demonstrated by recent voluntary guidelines issued from the White House to the G7, from Bletchley Park to Seoul. But do AI companies actually make good on their commitments? We score 16 companies based on their publicly disclosed behavior by developing a detailed rubric based on their eight voluntary commitments to the White House in 2023. We find significant heterogeneity: while the highest-scoring company (OpenAI) scores 83.3% overall on our rubric, the average score across all companies is just 54%. The companies demonstrate systemically poor performance on their commitment to model weight security, with an average score of 17%: 11 of the 16 companies receive 0% for this commitment. Our analysis highlights a clear structural shortcoming that future AI governance initiatives should correct: when companies make public commitments, they should proactively disclose how they meet their commitments to provide accountability, and these disclosures should be verifiable. To advance policymaking on corporate AI governance, we provide three directed recommendations that address underspecified commitments, the role of complex AI supply chains, and public transparency that could be incorporated into AI governance initiatives worldwide.

Introduction

The growing importance of artificial intelligence (AI) has rapidly catalyzed global policymaking efforts. Policymaking related to AI addresses many concerns including open innovation, market concentration, risk management, corporate governance, and geopolitics. Since 2023, many AI policy efforts have centered on the interplay between corporate governance, given that prominent AI systems are developed by the world's most powerful companies, and risk reduction, due to the breadth of potential harms associated with AI systems.

The approach to global AI policy varies significantly across jurisdictions. A key differentiator among jurisdictions that regulate AI companies is whether a policy imposes mandatory or voluntary obligations on companies. Some jurisdictions have enacted mandatory requirements via legislative or executive action, such as the EU AI Act and the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence respectively. However, much of global AI policy centers on voluntary actions taken by major companies in line with recommendations by government bodies. Key examples include the NIST AI Risk Management Framework, the 2023 White House Voluntary Commitments on AI, the G7 International Code of Conduct, Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, the 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse, and the Frontier AI Safety Commitments secured at 2024 AI Seoul Summit. Voluntary measures offer flexibility in that they can allow companies to pilot different approaches to meeting commitments, optimize for objectives other than minimizing legal risk associated with regulatory compliance, and harmonize approaches across jurisdictions despite different legal and political systems.

But policy initiatives that rely on companies to voluntarily take action have a number of pitfalls. Voluntary measures do not come with penalties for noncompliance, meaning that companies may choose to not participate, claim they are participating but not implement the government's recommendations, opt for partial implementation, or implement recommendations in ways that are opaque or not verifiable. Well-intentioned companies may have difficulty complying because voluntary measures are less likely to move markets and reorganize supply chains, meaning that measures requiring coordinated action may be less likely to succeed if voluntary. Voluntary measures often lack any mechanism for monitoring implementation, presenting a potential loophole for noncompliance (Aragón-Correa, Marcus, and Vogel 2020). A company's public commitment that it will adhere to voluntary measures can give the illusion that the company is taking significant action to responsibly develop and deploy AI systems while it does not in fact make any changes.

To understand the impact and efficacy of voluntary commitments, we conduct the first comprehensive analysis of the first major commitments to governments made by top AI companies. [1] In 2023, the White House secured voluntary commitments from 15 AI companies.[2] In announcing the commitments, the White House described their purpose as

[1]This version of the paper has been abridged for the AIES format. Please find the full version on arXiv instead: https://arxiv.org/abs/2508.08345.

[2]The commitments were secured in three phrases: (i) Amazon,

follows: "These commitments, which the companies have chosen to undertake immediately, underscore three principles that must be fundamental to the future of AI – safety, security, and trust – and mark a critical step toward developing responsible AI. As the pace of innovation continues to accelerate, the Biden-Harris Administration will continue to remind these companies of their responsibilities and take decisive action to keep Americans safe." Although the Biden Administration's AI Executive Order was later rescinded, the voluntary commitments secured from companies were not undone.

To reason about the companies and their behavior, we score companies based on how their public actions address their stated commitments. We design a scoring rubric that transforms the eight commitments specified by the White House on product safety, system security, and public trust into 30 indicators. Our rubric provides concrete and decidable criteria for determining if a company has satisfied its commitment. To score the 16 companies that signed the 2023 White House Voluntary Commitments on AI, for each of the 480 (indicator, company) pairs, we gather relevant public information through December 31, 2024, assign a score, and provide evidence for our decision.

By compiling information about company practices and interpreting it via quantitative scores, we provide evidence for three key findings. First, the scores demonstrate significant heterogeneity in companies' actions: the top-scoring company (OpenAI) scores 83% on our rubric, whereas the bottom-scoring company (Apple) scores 13%. Of the eight commitments, there are six commitments where at least one company scores 100%; at the same time, there are five commitments where at least one company scores 0%. Second, company-level scores demonstrate two clear, and interconnected, correlations: members of the Frontier Model Forum and earlier signatories tend to score higher. The six highest scoring companies are the six members of the Frontier Model Forum (OpenAI, Anthropic, Google, Microsoft, Meta, Amazon) and each score at least 60%. Third, model weight security is a commitment with distinctively poor performance: companies score on average 17%. 11 companies score 0% on this commitment (Adobe, Apple, Cohere, IBM, Inflection, Meta, Nvidia, Palantir, Scale AI, Salesforce, Stability AI).

Beyond providing empirical insight into the relationship between company practices and stated commitments, our work reveals a key design flaw in the 2023 White House Voluntary Commitments on AI: companies made public commitments to the White House, but no mechanism was created to monitor implementation or provide the public with information about implementation. To improve the design of future voluntary commitments related to corporate AI governance, we provide three recommendations to policymakers.[3]

1. **Commitments should be precise and specific.** The wording of the 2023 White House Voluntary Commitments is often vague, leading to significant ambiguity over the intent of a commitment and the steps required to satisfy a commitment. Commitments should be precise, specifying (i) what is the specific goal and (ii) what evidence is sufficient or satisfactory to indicate completion.
2. **Commitments should be targeted.** Since the same commitments are directed towards companies with different business models and roles in the AI supply chain, some commitments appear inappropriate for some companies (e.g. increased cybersecurity around model weights for companies that (largely) do not develop models). In contrast, commitments should be tailored to either (i) specific companies (e.g. if they operate across several levels of the supply chain) or (ii) a specific layer of the supply chain, clearly designating which companies belong to that layer.
3. **Commitments should enable public verification.** Though the 2023 White House Voluntary Commitments on AI were issued more than two years ago, the actions that companies have taken in order to fulfill their stated commitments remains highly uncertain based on public information. Given that these commitments are made publicly, we recommend that commitments include accountability measures (e.g. companies publish a transparency report six months after making commitments to indicate what actions they took for each commitment), especially to clarify whether companies changed their actions relative to what they may have done absent making such commitments.

The 2023 White House Voluntary Commitments on AI

Context. In 2023, the White House secured eight voluntary commitments with 15 leading AI companies: they are "commitments that companies are making to promote the safe, secure, and transparent development and use of generative AI (foundation) model technology" (White House 2023). At a high level, these commitments indicate that companies who are signatories will uphold three duties: (i) ensure their products are safe before public release, (ii) implement security practices for their AI models and systems, and (iii) earn public trust through responsible AI development. The commitments stated that companies intend to follow these commitments, alongside existing laws, until regulations that cover the same issues come into force.

Scope. In the initial July 2023 round of voluntary commitments, signed by seven companies at the time, the commitments were scoped to "generative models that are overall more powerful than the current industry frontier (e.g. models that are overall more powerful than any currently released models, including GPT-4, Claude 2, PaLM 2, Titan and, in the case of image generation, DALL-E 2)". When the White House announced in September 2023 that eight additional companies had signed, it modified the scope of the commitments to "generative models that are overall more powerful than the current most advanced model produced by

Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI committed in July 2023; (ii) Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability AI committed in September 2023; and (iii) Apple committed in July 2024, following the launch of its Apple Intelligence product.

[3]Unlike the 2023 White House Voluntary Commitments on AI, we find the 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse adopt some of our recommendations.

the company making the commitment".

Commitments. The first commitment is to conduct internal and external red-teaming of models or systems, focusing on risks including chemical, biological, radiological, and nuclear threats, cyber capabilities, autonomous system control, societal risks, and broader national security concerns. The second commitment addresses information sharing with different parties (e.g. other companies and governments) around trust and safety concerns, dangerous or emergent capabilities, and attempts to circumvent safeguards. Together, these commitments address the topic of product safety.

The next two commitments address system security. The third commitment covers the protection of proprietary and unreleased model weights through model-level cybersecurity, safeguards against insider threats, and personnel-level restricted access. Building on these company-internal practices, the fourth commitment encourages external discovery of vulnerabilities via bounties for third-party reporting.

The final four commitments collectively address public trust. These span commitments around content provenance methods and standards (commitment five), public reporting on capabilities and safety (commitment six), research on societal risks including empowering internal trust and safety teams (commitment seven), and prioritizing progress on society's greatest challenges as well as student, worker, and citizen engagement (commitment eight).

Scoring Methodology

To score companies, we define 30 indicators, gather public information on these indicators for each company, and use this information to support our score. Our methodology is inspired by the 2023 Foundation Model Transparency Index (Bommasani et al. 2023a).

Indicators

The White House commitments (White House 2023) are written as a combination of specific actions expected of companies and a more generic description of why these actions advance the public interest. As written, the commitments do not provide decidable criteria for determining whether a company's actions are sufficient to state that they fulfilled the commitment. Therefore, we define concrete *indicators* that transform each high-level commitment into more specific, decidable criteria that we use to score companies. To maximize fidelity with the voluntary commitments, each indicator is a verbatim excerpt from the commitments. The reference text for each is in Appendix A. Since the commitments vary in scope and content, we map each commitment to multiple indicators based on its wording. The resulting mapping (see Figure 1) yields 2–7 binary indicators per commitment and 30 indicators overall.

As an example, consider the seventh voluntary commitment on public trust, which is entitled "Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy". The commitment states: "Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized." We map this commitment to four indicators: (i) does the company empower its trust and safety teams? (ii) does the company advance AI safety research? (iii) does the company take steps to advance privacy? and (iv) does the company take steps to protect children?

We score each (company, indicator) pair on a binary basis. A score of 1 signifies that our search process surfaced publicly available documentation from the company that is sufficient to demonstrate that the company satisfied the portion of the 2023 White House Voluntary Commitments on AI captured by that indicator. A score of 0 signifies that our search process did not surface such documentation, whether because the documents identified did not contain sufficient evidence to demonstrate the commitment was fulfilled or because no relevant documents were found through our search process.

We construct binary indicators for several reasons. First, our aim is to break the commitments down into distinct, decidable chunks that can be used to assess whether or not there is sufficient evidence that a specific sub-part of a commitment was or was not fulfilled. Second, producing narrower criteria for scoring reduces subjectivity in assigning initial scores. Third, binary indicators simplify the scoring process by allowing scorers to focus on the sharp distinction between 0 and 1 point for each indicator (Bommasani et al. 2023a).

We acknowledge that binary indicators are potentially reductive, leaving out valuable information that can be captured by more complex scoring schemes. At the same time, a greater number of smaller, binary indicators can be aggregated to produce more complex scoring schemes, and the information we release associated with our scores could be used to produce alternate scores using different criteria.

Information Gathering

To score companies, we used public information released by the companies with no additional third-party sources. In doing so, we highlight that companies, with the exception of commitment six on public reporting, did not commit to making such information publicly available. It is therefore possible that companies do satisfy their voluntary commitments but do not provide any public evidence of implementation. Given the high-profile and public nature of these commitments and companies' statements in support of public transparency (Bommasani et al. 2023a), we believe it is appropriate to assess companies based on their public disclosures.

Nevertheless, companies may be motivated by values and interests other than public transparency. For example, concerns regarding security may lead companies to not disclose information on their model-weight security practices and insider threat detection programs. In some cases, companies may lack the authority to unilaterally disclose information related to their commitments, including information that has been shared with governments and/or other companies.[4] We emphasize the opportunity for Pareto improvement: compa-

[4]Potential motivations for a lack of transparency on matters like research into how frontier AI systems can help meet society's greatest challenges may be less well grounded, though absolute transparency could conflict with commercial interests.

Commitment	Indicator
Red-teaming	Internal red-teaming
	External red-teaming
	Red teaming coverage of risks
Information Sharing	Information sharing with companies
	Information sharing with government
	Forum or mechanism for information sharing
	Forum or mechanism shares information on risks
Model weight security	Model weight cybersecurity practices
	Insider threat detection program
	Limiting weight-level access to relevant personnel
Third-party reporting	Establish bounties, contests, or prizes
	Include AI systems in their existing bug bounty programs
Watermarking/Provenance	Robust provenance or watermarking for audio
	Robust provenance or watermarking for visual content
	Develop tools or APIs to determine if a particular piece of content was created within their tools
	Work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI
Public reporting	Report capabilities
	Report limitations
	Report domains of appropriate use
	Report domains of inappropriate use
	Report safety evaluations
	Report on societal risks
	Report on adversarial testing used to determine appropriateness of deployment
Societal risk research	Empower trust and safety teams
	Advance AI safety research
	Advance privacy
	Protect children
Address society's greatest challenges	Support research and development of frontier AI systems that can help meet society's greatest challenges, such as climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats.
	Support initiatives that foster the education and training of students and workers to prosper from the benefits of AI
	Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology.

Table 1: Indicators. Table of the 30 indicators we use to score companies.

nies likely can provide some additional information on their conduct to the public without any tradeoff with their financial, reputational, or security interests.

We score companies based on information we gathered by December 31, 2024—our scores do not reflect new information that was made available thereafter or models that have been released since.[5] We use information that is deliberately and directly disclosed by the company—other sources such as leaked information, media reporting, or external analysis is not used. These decisions contribute to greater fairness when assessing companies and comparing their scores, as companies themselves control their scores by deciding what information to publish about their behavior.

We gathered information in a three stage process. First, we collected key reference documents for each company that describe their practices in relation to their generative AI models, systems, and products. These documents include (a) external-facing resources such as blog posts, press releases, and transparency reports, (b) resources useful for the research community such as research papers, technical reports, model cards, documentation for developers, and bug bounties, as well as (c) product policies and safety frameworks. These documents were identified through an initial review of publicly available materials for each company and then selected based on their relevance to the commitments. We prioritized materials that explicitly address how companies assess, mitigate, or communicate risks associated with their generative AI systems, as well as those that provide insight into internal governance structures or external accountability mechanisms.

Second, we searched through these documents and produced additional resources by creating a search script and using a language model for standardized, automated search. For each (company, indicator) pair, we use the script to better narrow our search. We query the Perplexity API with the following search string: "What has {COMPANY_NAME} done since the beginning of 2023 that might fit under: {INDICATOR TEXT}? Make sure to return links used to find this information. Keep it concise and make sure to return all links with no information from before 2023."[6] For each link returned in the Perplexity response, we reviewed the source document for relevance. We note that Perplexity was used only to augment our information gathering process, not as a substitute for our manual search.

Third, we compiled the sources resulting from the first two steps for every (company, indicator) pair as the basis for making scoring decisions.[7] While these compiled sources are not exhaustive—in significant part because companies often deprecate documents on their websites, bury important documentation several layers deep, or fail to adequately summarize their actions to fulfill public commitments—we reviewed hundreds of documents as part of this process.

Scoring

For each of the 16 companies, we use the information gathered from the above process to produce initial scores for each of the 30 indicators.

As we scored indicators and identified disagreements among scorers, we iteratively developed specific and measurable criteria to evaluate fulfillment of each indicator, requiring in every instance that evidence be publicly verifiable. These criteria reflect our interpretation of whether company actions align with the goals underlying the commitments, while remaining grounded in their language and scope.

For instance, to assess if the company empowers its trust and safety team, we consider whether (1) the company explicitly identifies such a team and (2) the company's documentation indicates it adequately resources the team and/or provides it the authority to address potential risks. The criteria for every indicator can be found in Appendix D.

Two authors of this paper each independently assigned an initial score for every one of the 480 (company, indicator) pairs. Both authors provided a source and a quote to justify each score. In the event of disagreement on a particular score, all of the authors of this work discussed, coming to agreement in assigning the final score.

The agreement rate was 75.6% ($\frac{363}{480}$), reflecting substantial agreement. However, the ambiguity in the wording of the commitments and how they apply to each specific company was a core source of initial disagreement, as was the variation in the level of detail across companies' public documentation. We release the final score for all 480 (company, indicator) pairs along with a justification for the score and associated reference(s) to public materials.

In the event that an indicator is related to a specific model or system (e.g. whether the company implements model-weight cybersecurity practices), we score the company based on its flagship foundation model or system as of December 31, 2024.[8] We choose the flagship foundation model as an object of analysis because the September 2023 version of the commitments focus on the capabilities of the "most advanced model" for each company, while the July 2023 version explicitly named several companies' flagship models. In addition, many companies make their flagship foundation models (or derivatives) central to the bulk of their AI-based products and services due to their enhanced capabilities. The mapping from companies to flagship models is provided in Appendix C. We acknowledge that other models and systems beyond the flagship models we consider may also fall in scope of the commitments.

Results

To organize our analysis, we apply three lenses: (i) an overall company-level view, (ii) a commitment-level view, and (iii) a disaggregated indicator-level view.

[5]Since some companies signed onto the commitments at different times, with Apple being a notable outlier in 2024 (compared to the other 15 companies in 2023), companies had varying amounts of time between their commitment and our scoring.

[6]We considered various search APIs (including those from OpenAI, Anthropic, and Google) and prompts, eventually finding the Perplexity API performed best at surfacing new relevant documents.

[7]The search scripts and compiled sources are released publicly under an MIT license at https://github.com/rishibommasani/whvc.

[8]The flagship model is defined as in Bommasani et al. (2023c): "the foundation model that is most salient and/or capable from the developer based on our judgment, which is directly informed by the company's public description of the model."

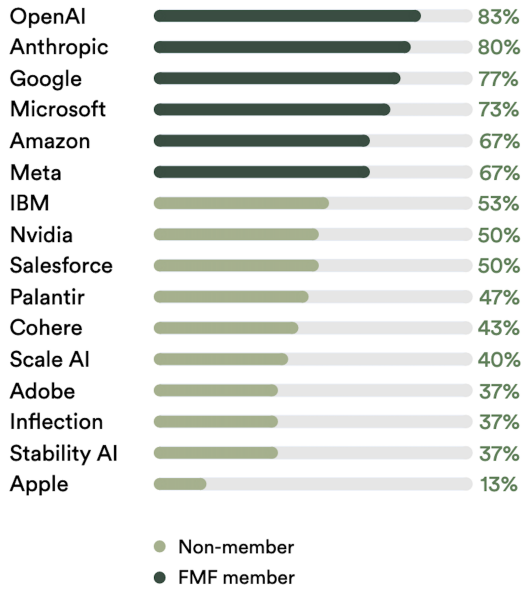

Figure 1: Aggregate scores by company. The score for each company, stratified by whether the company belongs to the Frontier Model Forum (FMF) as of December 31, 2024.

In Figure 1, we report the aggregate score as a percentage for each company across all the indicators. The mean and median are 53.3% and 50.0% respectively, with a standard deviation of 19.5%. The range is 70.0% between the highest scoring company, OpenAI, at 83.3% and the lowest scoring, Apple, at 13.3%. While OpenAI satisfies 25 of the 30 indicators,[9] no company has a perfect score despite making these commitments to the White House over two years ago.

Significant variation in companies' scores. There is notable variation in how companies perform, with companies clustering into three distinct groups. Four companies score at least one standard deviation above the mean: OpenAI (83.3%), Anthropic (80.0%), Google (76.7%), and Microsoft (73.3%). The majority of companies fall within one standard deviation of the mean: Amazon (66.7%), Meta (66.7%), IBM (53.3%), Nvidia (50.0%), Salesforce (50.0%), Adobe (36.7%), Cohere (43.4%), Palantir (36.7%), Inflection (36.7%), Stability AI (36.7%), Scale AI (36.7%). The only company that scores at least one standard deviation below the mean is Apple (13.3%).

Company-Level Results

Frontier Model Forum members consistently score higher. Strikingly, the company-level scores clearly separate based on membership in the Frontier Model Forum (FMF). The Frontier Model Forum is a non-profit industry association dedicated to advancing the safe development and deployment of frontier AI systems (Frontier Model Forum 2025a). Anthropic, Google, Microsoft, and OpenAI became the four founding members in July 2023, with Amazon and Meta joining in May 2024. The six highest scoring companies, which all score at least 66%, are the six FMF members. Their mean score is 74.4% with a standard deviation of 6.9%. In contrast, the 10 other companies all score at or below 60% with a mean of 40.7% and a standard deviation of 11.4%. It is notable that FMF, in consultation with its members, has published technical reports intended in part to facilitate compliance with voluntary commitments (Frontier Model Forum 2025b). FMF states that its technical reports aim to "examine how [Frontier AI] frameworks can be implemented effectively" and acknowledges that such frameworks are the core component of the Frontier AI Safety Commitments at 2024 AI Seoul Summit (Frontier Model Forum 2025c).

Earlier signatories generally score higher. The 16 companies signed onto the voluntary commitments in three phases of participation: seven companies in July 2023 (Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI), eight companies in September 2023 (Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, Stability AI) and one company in July 2024 (Apple). We find company-level scores are clearly correlated with the timing of signature. The first cohort has a mean of 69.0% with a standard deviation of 15.6%, while the second cohort has a mean of 44.6% with a standard deviation of 6.4%. It is possible this disparity reflects additional time the first cohort had before our scoring to publish documentation, but given that both cohorts had over 15 months prior to scoring, we hypothesize that the first cohort's business practices better align with the commitments.

Commitment-Level Results

High scores for content provenance due to non-applicability. Based on the average per-commitment score for each company, the clear highest-scoring commitment is for (audiovisual) watermarking and provenance at 92.2%. 14 companies receive 100% for this commitment.[10] In many cases, however, companies satisfy the associated indicators vacuously, because they do not develop audio or visual models, which are the subject of the provenance commitment. Still, of the 8 companies that develop models with these output modalities, the average is 83.9%, which exceeds that of all other commitments. In particular, many of these companies follow industry standards associated with the Coalition for Content Provenance and Authenticity (C2PA) and their Content Credentials; 6 of the 8 companies are steering committee members of C2PA, while Apple and Stability AI are unaffiliated. In contrast to the high scores for this commitment from most companies, Apple is the sole outlier as a company with audio and visual models that scores 0% on these indicators.

[9]The indicators that OpenAI does not satisfy are: "Insider threat detection program", "Report limitations", "Report domains of appropriate use", "Empower trust and safety teams", and "Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology".

[10]These companies are: Adobe, Amazon, Anthropic, Cohere, Google, IBM, Inflection, Meta, Microsoft, Nvidia, OpenAI, Palantir, Scale AI, Salesforce.

Low scores for model-weight security in spite of global emphasis. Based on the averages, the lowest scoring commitment is on model-weight security at 22.9%. Eleven companies score 0% on this commitment, and none receives a full marks. The high-scoring companies are OpenAI, Anthropic, and Microsoft at 66.7%. Anthropic is the only company that indicates the existence of an insider risk program as part of its security standard. OpenAI and Microsoft, on the other hand, both state they create a secure research environment dedicated to model security and implement an access control protocol. While transparency around model-weight security practices is valuable, we acknowledge that maximal transparency about security practices for model weights could undermine that very security. However, the fact that every indicator is met by at least one company suggests that Pareto improvements are possible in how other companies navigate the transparency-security trade-off. We emphasize the current results are particularly concerning given how model-weight security remains a clear challenge (Nevo et al. 2024) and features in many global AI policies (e.g. the sixth commitment of the G7 International Code of Conduct (Group of Seven 2023), Section 3.1 of the US AI Safety Institute guidance on Managing Misuse Risk for Dual-Use Foundation Models (U.S. AI Safety Institute 2024), Section 4 of US Executive Order 14141 on Advancing United States Leadership in Artificial Intelligence Infrastructure (The Executive Office of the President 2025)).

Low scores for third-party reporting align with concerns of chilling effects on third-party research. Alongside model-weight security, third-party reporting is another low-scoring commitment at 34.4%. Eight companies score 0% on this commitment. These low scores are especially surprising because the commitment is focused on providing bounties for reporting, and there are natural incentives for companies to make these bounties transparent to maximize external reporting. Our finding aligns with those of Longpre et al. (2024), who find that current company policies around AI-related bug bounties and protections for third-party research are unclear and uneven. In particular, they argue that companies' policies suppress third-party reporting—given that researchers may be concerned with legal reprisal absent safe harbor (e.g. for responsible penetration testing)—instead of being supportive of such research, as required by this commitment.

Indicator-Level Results

Extreme indicator-level scores align with commitment-level scores. On average, each indicator is awarded to 8.5 of the 16 companies with a standard deviation of 4.9. Seven indicators are satisfied by at least 14 companies (one standard deviation above the mean): four belong to the highest-scoring commitment on content provenance, two belong to the commitment on public reporting. These are "Report capabilities", which is satisfied by every company and is clearly incentivized by market forces, and "Report domains of inappropriate use", which is satisfied by every company except for Apple. The remaining indicator is to "Establish or join a forum or mechanisms for information sharing", which all companies receive on the basis of their membership in the US AI Safety Institute Consortium.

In contrast, five indicators are scored by at most three companies (one standard deviation below the mean): one is "Insider threat detection program" under the low-scoring model weight security commitment. The other four are (i) "Information sharing with government", which only OpenAI and Anthropic satisfy by establishing memoranda of understanding with the US AI Safety Institute, (ii) "Empower trust and safety teams", which Google and Inflection satisfy by integrating trust and safety assessments into the model pre-launch processes and authorizing their teams to use a full range of tools to block malicious actors, (iii) "Red teaming coverage of risks", which OpenAI and Anthropic satisfied by conducting red-teaming exercises that address all the risk areas specified in the commitment, and (iv) "Support initiatives that help citizens understand the nature, capabilities, limitations, and impact of the technology", which none of the companies satisfied.

Indicator-level analysis reveals substantial heterogeneity in information sharing. The information sharing commitment spans four indicators: information sharing with other companies (56.3%), information sharing with governments (12.5%), forum or mechanism for information sharing (100%), and forum or mechanism that discloses information on risks (43.8%). While every company satisfies the indicator for a forum or mechanism for information sharing due to participation in the US AI Safety Institute Consortium, we do not automatically award the further point for sharing information on risks because it is not clear that this occurs in the Consortium. Only seven companies are awarded this indicator, largely based on Frontier Model Forum membership.

Further, while the US AI Safety Institute Consortium was established by a governmental body in the National Institute of Standards and Technology, we do not automatically designate it as a means for information sharing with the government because our standard is that shared information should be non-public and do we not find evidence that companies share such information with the government through the Consortium. As a result, only OpenAI and Anthropic score this indicator on the basis of their memoranda of understanding with the US AI Safety Institute, which permit US AISI to directly access their models to perform risk assessments. While companies do interface with the government in other ways—such as procurement of companies' AI systems, Congressional testimony from executives, and enforcement investigations into company practices—these are insufficient to satisfy this indicator.

Certain indicators are overly vague, complicating consistent interpretation and meaningful implementation. While every indicator is only partially specified by the White House in its three-page document describing the voluntary commitments, some indicators are especially vague. The clearest example is commitment seven, where "Companies commit generally to empowering trust and safety teams, advancing AI safety research, advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized". All four of the resulting

indicators are exceptionally broad and difficult to judge: what constitutes satisfactory privacy advancement or protection of children? Even less clear is how these commitments are meant to relate with company practices on AI: for example, moderating the generation of child sexual abuse material and monitoring the use of language models by young children may both serve to protect children in very different senses.

Without concrete definitions to delineate what companies should do, companies and the public are highly unlikely to interpret the commitments in the same way. In scoring these commitments, we chose to award points for constructive steps that met what we considered the minimum viable standard for public accountability and the maximally defensible standard absent greater clarification from the White House. Even if companies simultaneously demonstrated contradictory behavior, we credited them for taking steps aligned with the commitments (in order to establish a consistent baseline on company adherence). For example, Meta received the point for "Protecting children" for its partnership with Thorn and the National Center for Missing & Exploited Children, although the end-to-end encryption on its platforms prevents the detection of child sexual exploitation. Taken together, the vagueness in how the commitments are articulated and the uncertainty regarding how to assess a company's practices in totality lead us to question whether such high-level commitments are meaningful.

Related Work

To contextualize our work, we discuss prior work that assesses major AI companies based on their public conduct and discuss other voluntary commitments.

Assessments of AI Companies for 2023 WHVC

Beyond our work, the most comprehensive analysis of company practices in relation to these commitments was conducted as part of a MIT Technology Review article published on the one-year anniversary of the commitments (Heikkilä 2024). As part of this work, Heikkilä (2024) contacted the seven initial signatories and received responses from six of these companies, excluding Inflection, on how they addressed each of the commitments; external researchers also provided commentary. Overall, the work found evidence that companies had taken steps to implement some technical model-level interventions(e.g. red-teaming and watermarking) and made investments in safety research. However, less evidence was found related to progress on information sharing, third-party reporting and public reporting.

Heikkilä (2024) indicates that no comprehensive evaluation had been performed of the commitments, company practices, or their relationship. In light of this, our work not only provides a comprehensive assessment, but also introduces a concrete scoring system that yields quantitative findings. In general, our findings largely agree with those of Heikkilä (2024) and Roose (2023), with the main difference being the depth and specificity of our results, though we highlight that our scores are based on public information from companies whereas the prior work only considered the brief responses companies provided to journalists. Further, our work expands the focus to the full set of 16 companies, rather than just the initial seven, which enables us to identify clear disparities between the initial signatories and the remaining signatories.

Assessments of AI Companies

As technology companies have grown in importance and become some of the world's most powerful entities, a multidisciplinary body of literature has emerged to assess these companies with a variety of methods. In the space of quantitative assessments, several works have introduced scoring approaches either in the form of one-off analyses, akin to this work, or sustained indices, which score the same companies on a recurring cadence. As an illustrative example, we highlight the Corporate Accountability Index that is maintained by Ranking Digital Rights (RDR), which has scored telecommunication and technology companies since 2015 for how they "respect users' fundamental rights, and on the mechanisms they have in place to ensure those promises are kept" (Dheere et al. 2020). Kogen (2024) analyzed the 2018 Index and showed, by reviewing internal RDR documents and interviewing relevant stakeholders (e.g. representatives from 11 companies and 14 civil society groups), that it usefully communicated legible, newsworthy, and flexible information that empowered social movements.

Drawing upon this tradition, several recent works have employed and developed similar scoring approaches for the assessment of AI companies (Bommasani et al. 2023b, 2024; Klyman 2024; Longpre et al. 2024; AI Lab Watch n.d.; hEigeartaigh et al. 2023; Barrett, Newman, and Nonnecke 2023; Jones n.d.). To our knowledge, Bommasani et al. (2023b) provided the first assessment of major AI companies by scoring them on a rubric based on the European Parliament's proposal for the EU AI Act. Based on the results, they made evidence-based recommendations aimed at (i) EU legislators on how the EU AI Act should be updated during the legislative negotiation and (ii) companies on how they could modify their practices to better align with the proposed requirements. While some works similarly link scoring to specific governmental policies (e.g. Barrett, Newman, and Nonnecke (2023) assess companies in relation to NIST's AI Risk Management Framework, hEigeartaigh et al. (2023) score in relation to the UK's recommendations), other works provide independent specification of the indicators or criteria of interest. The Foundation Model Transparency Index is an annual index that scores foundation model developers for their transparency across the supply chain with 100 indicators that span the resources used to build a model (e.g. data, compute), the properties of the model itself (e.g. capabilities, risks), and the use of the model in society (e.g. distribution, impact) (Bommasani et al. 2023a, 2024).

Cumulatively, these works all demonstrate a shared methodology of scoring companies with different approaches for sourcing the indicators, determining the scores, and theories of change for how the results and takeaways improve corporate governance and/or public policy. Many of these works also share two key findings with our work. While all of these works aim to increase public accountability, they all encounter limits due to the lack of transparency into company-internal practices. And, while the exact magnitudes

and details often differ, these works almost always find considerable heterogeneity in company practices. Together, they highlight the absence of clear norms, let alone more formal mechanisms, for ensuring public-facing transparency and standardizing industry-wide conduct.

Voluntary Commitments From Governments

Global AI policy reflects a broad constellation of efforts that spans long-standing policy in specific domains (e.g. applying hiring discrimination laws to algorithmic hiring), more recent policy for digital technologies (e.g. applying data protection laws to training data), and new policy for AI specifically (e.g. new laws to govern AI). While many jurisdictions face shared challenges, the overall global AI policy landscape reflects significant heterogeneity that indicates both region-specific considerations and idiosyncratic differences. In particular, when considering AI-specific policy, several jurisdictions currently employ voluntary approaches to corporate governance with the European Union's approach via the EU AI Act standing as a clear counter example. At this juncture, given many of these voluntary and/or mandatory policies are very recent, little evidence exists to empirically validate the strengths and/or weaknesses of these two top-level approaches.

As a result, we briefly survey some of the voluntary commitments and approaches taken elsewhere in the world to contextualize the approach taken in the 2023 White House Voluntary Commitments on AI. The U.S. NIST AI Risk Management Framework, as well as the associated profile on generative AI in particular, provides voluntary guidance to help organizations identify, assess, manage, and mitigate risks by emphasizing trustworthy AI principles such as fairness, transparency, accountability, security, and privacy (Tabassi 2023). The Canada Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems introduces voluntary commitments applicable to the responsible development and deployment of foundation models, such as accountability, safety, fairness, human oversight, and robustness, as well as for developers and managers of generative AI systems (ISED Canada 2023). The G7 International Code of Conduct for Organizations Developing Advanced AI Systems articulates 11 commitments that span data protection, risk management, technical standard, and transparency reporting: while companies not signed on in the same way they have done for certain national-level commitments, these commitments may serve as the basis for global agreement (Group of Seven 2023). Most recently, the Biden-Harris Administration secured voluntary commitments with AI model developers and data providers to prevent and mitigate the misuse of AI in creating and disseminating image-based sexual abuse content (White House 2024).

Beyond these standalone voluntary commitments, the ongoing series of international AI Summits have emerged as a key generative process for voluntary commitments as global policymakers work together to advance AI governance. Beginning with the U.K. AI Safety Summit in November 2023, the Bletchley Declaration was signed by 29 world governments to foster international cooperation on AI policy through an agenda centered on (i) "identifying AI safety risks of shared concern, building a shared scientific and evidence-based understanding of these risks, and sustaining that understanding . . . " as well as (ii) "building respective risk-based policies across our countries to ensure safety . . . alongside increased transparency by private actors developing frontier AI capabilities, appropriate evaluation metrics, tools for safety testing, and developing relevant public sector capability and scientific research". To advance this agenda, at the subsequent AI Seoul Summit in May 2024, 16 global companies (Amazon, Anthropic, Cohere, Google, G42, IBM, Inflection AI, Meta, Microsoft, Mistral AI, Naver, OpenAI, Samsung Electronics, Technology Innovation Institute, xAI, Zhipu.ai) signed onto the Frontier AI Safety Commitments. The associated eight commitments address three outcomes: (i) improved risk management practices, (ii) increased accountability for safe development and deployment and (iii) sufficient transparency to external stakeholders. Building on these efforts, the United States convened the growing global network of AI Safety Institutes in November 2024 for a working meeting on three high-priority topics (managing risks from synthetic content, testing foundation models, and conducting risk assessments for advanced AI systems) that articulated six principles for risk assessment (actionability, transparency, comprehensiveness, multi-stakeholder consideration, iterativity, and reproducibility).

Discussion

Our research into the voluntary commitments leads us to consider: (i) future-looking commitment and policy design, and (ii) current corporate practices and governance.

Commitments should be clearly worded. The White House Voluntary Commitments were first announced with a fact sheet and an accompanying three-page document. While these documents are likely intended for public consumption and, therefore, provide generic high-level description, they are ambiguous. In particular, some commitments are vague in terms of their intent (e.g. language such as "protect children"), especially when targeted at companies with large footprints and many roles in the AI supply chain. Further, all commitments lack conditions for what constitutes satisfactory conduct. While voluntary approaches permit flexibility to avoid being overly prescriptive or burdensome, these goals are achievable while still communicating about what is desired, especially for actions that can vary greatly in magnitude (e.g. how much internal or external red-teaming is desired?). We recommend that commitments be precisely worded so that they articulate specific goals along with what constitutes sufficient evidence of completion. Practically, these lower-level details may need to be split out into appendices or supporting documents, but the goal of broad intelligibility for the public need not be at odds with meaningful precision for deeply engaged stakeholders.

Commitments should be clearly targeted. The voluntary commitments, across their three phases of signing, specify essentially the same commitments for all 16 signatories. However, these companies occupy significantly different positions in the AI ecosystem: they differ in their business models, their set of roles in the supply chain, and how

their AI-related practices mediate public outcomes. Given their uniform treatment under the commitments, some commitments generally made little sense for certain companies (e.g. increased cybersecurity around model weights for companies that (largely) do not develop models). While these commitments could have future-facing utility, we ultimately are skeptical of this one-size-fits-all approach, especially given our empirical findings that massive technology companies may take positive action in one part of their business practice while regressing in another. We recommend that commitments either be tailored for each company or, when trying to standardize across companies, be tailored to a specific supply chain role. The 2024 White House Voluntary Commitments to Combat Image-Based Sexual Abuse adopts this approach: for example, Meta and Microsoft have differentiated obligations that reflect how they operate different platforms downstream that contribute to the distribution of this imagery.

Commitments should enable public verification. The voluntary commitments, except for commitment six on public reporting, specify no means for the public to understand or verify how companies took action to realize their commitments. Empirically, our entire analysis and that of Heikkilä (2024) make clear that public insight is limited, even given more than a year has elapsed since the commitments were first made. This directly contradicts one of the three stated goals of the voluntary commitments, which is to increase public trust. Moving forward, commitments could be accompanied by accountability mechanisms (e.g. a standardized transparency report that articulates how specific company actions address specific commitments) to address the clear gap we observe. We recommend that public commitments, especially those made between very high profile institutions like the U.S. federal government and major AI companies, require periodic public transparency.

Concerning practices. Beyond the specific practices we score, we highlight that some companies have released materials or otherwise discussed their conduct in relation to the commitments. These companies include Amazon (Philomin 2024), Anthropic (Anthropic 2024), Google (Google n.d.), Meta (Meta 2023), Microsoft (Microsoft 2023), OpenAI (OpenAI 2024), Inflection (Inflection AI 2023), and Salesforce (Salesforce 2024). In reviewing these references, we at times disagreed with the company's claims that their conduct satisfactorily addresses the voluntary commitments. For example, Meta claims to have fulfilled the commitment on information sharing by publicly releasing artifacts about their models' capabilities and limitations. While these artifacts earned them points on public reporting, we only awarded points to companies for information sharing beyond public disclosure. Separately, Salesforce credits themselves for incentivizing third-party discovery through their bug bounty program to prevent AI-powered cyber threats. However, Salesforce does not specify that their AI systems are covered under the scope of this program, and therefore we did not award them the point on third-party reporting. In part, this reflects that these statements are often simultaneously high-level (e.g. "we're prioritizing cybersecurity safeguards to protect proprietary and unreleased models and we're participating in industry- wide events to support broader protections...") and are made without accompanying proof.

These statements compound the issues we raise on commitment design. If companies not only do not demonstrate how they addresses public commitments, but also broadly claim they satisfied the commitments based on their unilateral judgment, then the overall integrity of the commitments is further compromised. In turn, this further substantiates our recommendation for why standardized and timely reporting in response to public commitments is especially vital for these commitments to meaningfully advance corporate governance.

Promising practices. As a positive demonstration of how companies can communicate about their commitments, we point to the webpage Anthropic published on tracking their progress.[11] On the page, Anthropic enumerates every commitment they have made and how they map to actions they have taken. In particular, such a page also clarifies how overlapping commitments (e.g. commitments to conduct internal and external risk assessment that overlap across the White House Voluntary Commitments, the G7 International Code of Conduct, and the Frontier AI Safety Commitments) are streamlined by global companies operating in many jurisdictions. While this does not imply whether or not Anthropic meets our per-indicator standard, nor any standard the White House envisioned, it claries how Anthropic sees the correspondence between their actions and their commitments. All major AI companies could implement a similar approach to track how companies' internal practices and external commitments evolve.

Conclusion

We present the first comprehensive analysis of how leading AI companies have implemented major voluntary commitments made to governments. Our findings reveal substantial variation in implementation, with scores ranging from 83.3% for highest-scoring company to just 13.3% for the lowest. High-scoring companies tended to be early signatories and members of the Frontier Model Forum. However, overall performance was particularly weak in model weight security and third-party reporting. These findings underscore the need for future voluntary commitments to be clearly defined, appropriately targeted to specific roles of companies, and supported by mechanisms for public verification in order to meaningfully influence corporate behavior.

Acknowledgments

We thank Melissa Heikkilä, Miles Brundage, Miranda Bogen, Sayash Kapoor, Shayne Longpre, Percy Liang, Daniel E. Ho, and Daniel Zhang for discussions on this topic. RB is funded by the Stanford Lieberman Fellowship. This work is entirely unrelated to the involvement of RB in the EU AI Act Codes of Practice. KK completed this work at Stanford prior to his work in government.

[11] See https://www.anthropic.com/voluntary-commitments.

References

AI Lab Watch. n.d. Commitments. https://ailabwatch.org/resources/commitments/.

Anthropic. 2024. Tracking Voluntary Commitments. https://www.anthropic.com/voluntary-commitments.

Aragón-Correa, J. A.; Marcus, A. A.; and Vogel, D. 2020. The Effects of Mandatory and Voluntary Regulatory Pressures on Firms' Environmental Strategies: A Review and Recommendations for Future Research. *Academy of Management Annals*.

Barrett, A.; Newman, J.; and Nonnecke, B. 2023. AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models. https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf#page=83.

Bommasani, R.; Klyman, K.; Kapoor, S.; Longpre, S.; Xiong, B.; Maslej, N.; and Liang, P. 2024. The Foundation Model Transparency Index v1.1: May 2024. arXiv:2407.12929.

Bommasani, R.; Klyman, K.; Longpre, S.; Kapoor, S.; Maslej, N.; Xiong, B.; Zhang, D.; and Liang, P. 2023a. The Foundation Model Transparency Index. *ArXiv*, abs/2310.12941.

Bommasani, R.; Klyman, K.; Zhang, D.; and Liang, P. 2023b. Do Foundation Model Providers Comply with the Draft EU AI Act?

Bommasani, R.; Zhang, D.; Lee, T.; and Liang, P. 2023c. Improving Transparency in AI Language Models: A Holistic Evaluation. *Foundation Model Issue Brief Series*.

Dheere, J.; MacKinnon, R.; Brouillette, A.; Biddle, E. R.; Gutermuth, L.; Maréchal, N.; Renieris, E. M.; Wessenauer, V.; Rydzak, J.; Sperling, I.; Abrougui, A.; Rogoff, Z.; Zhang, J.; Bhatia, A.; Ross, K.; and Walton, G. 2020. 2020 Ranking Digital Rights Corporate Accountability Index.

Frontier Model Forum. 2025a. About Us. https://www.frontiermodelforum.org/about-us/.

Frontier Model Forum. 2025b. Frontier Capability Assessments. https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/.

Frontier Model Forum. 2025c. Introducing the FMF's Technical Report Series on Frontier AI Frameworks. https://www.frontiermodelforum.org/updates/introducing-the-fmfs-technical-report-series-on-frontier-ai-safety-frameworks/.

Google. n.d. Fulfilling the Voluntary Industry Commitments on AI. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/whcommitments.pdf.

Group of Seven. 2023. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Syste. https://www.mofa.go.jp/files/100573473.pdf.

hEigeartaigh, S. O.; Lannquist, Y.; Marcoci, A.; Sevilla, J.; Ruiz, M. A. U.; Chaudhary, Y.; Schreier, T.; Stein-Perlman, Z.; and Ladish, J. 2023. Do companies' AI Safety Policies meet government best practice? https://www.lcfi.ac.uk/news-events/news/ai-safety-policies.

Heikkilä, M. 2024. AI companies promised to self-regulate one year ago. What's changed? https://www.technologyreview.com/2024/07/22/1095193/ai-companies-promised-the-white-house-to-self-regulate-one-year-ago-whats-changed/.

Inflection AI. 2023. The precautionary principle: partnering with the White House on AI safety. https://inflection.ai/blog/partnering-with-the-white-house-on-ai-safety.

ISED Canada. 2023. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems. https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

Jones, A. n.d. How are AI companies doing with their voluntary commitments on vulnerability reporting? https://adamjones.me/blog/ai-vulnerability-reporting/.

Klyman, K. 2024. Acceptable Use Policies for Foundation Models. arXiv:2409.09041.

Kogen, L. 2024. From Statistics to Stories: Indices and Indicators as Communication Tools for Social Change. *The International Journal of Press/Politics*, 29(4): 1090–1108.

Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Blili-Hamelin, B.; Huang, Y.; Skowron, A.; Yong, Z.-X.; Kotha, S.; Zeng, Y.; Shi, W.; Yang, X.; Southen, R.; Robey, A.; Chao, P.; Yang, D.; Jia, R.; Kang, D.; Pentland, S.; Narayanan, A.; Liang, P.; and Henderson, P. 2024. A Safe Harbor for AI Evaluation and Red Teaming. *ArXiv*, abs/2403.04893.

Meta. 2023. Overview of Meta AI safety policies prepared for the UK AI Safety Summit. https://transparency.meta.com/en-gb/policies/ai-safety-policies-for-safety-summit.

Microsoft. 2023. An update prepared for the UK AI Safety Summit. https://blogs.microsoft.com/on-the-issues/2023/10/26/microsofts-ai-safety-policies/.

Nevo, S.; Lahav, D.; Karpur, A.; Bar-On, Y.; Bradley, H.-A.; and Alstott, J. 2024. *Securing AI model weights: Preventing theft and misuse of frontier models*. 1. Rand Corporation.

OpenAI. 2024. OpenAI Safety Update. https://openai.com/index/openai-safety-update/.

Philomin, V. 2024. A Progress Update on Our Commitment to Safe, Responsible Generative AI. https://aws.amazon.com/blogs/machine-learning/a-progress-update-on-our-commitment-to-safe-responsible-generative-ai/.

Roose, K. 2023. How Do the White House's A.I. Commitments Stack Up? https://www.nytimes.com/2023/07/22/technology/ai-regulation-white-house.html.

Salesforce. 2024. Tracking Our Progress on the White House Voluntary AI Commitments. https://www.salesforce.com/news/stories/voluntary-ai-commitments/.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

The Executive Office of the President. 2025. Advancing United States Leadership in Artificial Intelligence Infrastructure. Executive Order 14141, Federal Register Document 2025-01395, 90 FR 5469–5489.

U.S. AI Safety Institute. 2024. Managing Misuse Risk for Dual-Use Foundation Models. https://doi.org/10.6028/NIST.AI.800-1.ipd.

White House. 2023. Ensuring Safe, Secure, and Trustworthy AI. https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

White House. 2024. White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse. https://bidenwhitehouse.archives.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/.