July 25, 2023

Skadden

If you have any questions regarding the matters discussed in this memorandum, please contact the following attorneys or call your regular Skadden contact.

Stuart D. Levi
Partner / New York
212.735.2750
stuart.levi@skadden.com

David A. Simon
Partner / Washington, D.C.
202.371.7120
david.simon@skadden.com

Shannon N. Morgan
Associate / New York
212.735.3711
shannon.morgan@skadden.com

This memorandum is provided by Skadden, Arps, Slate, Meagher & Flom LLP and its affiliates for educational and informational purposes only and is not intended and should not be construed as legal advice. This memorandum is considered advertising under applicable state laws.

One Manhattan West
New York, NY 10001
212.735.3000

1440 New York Ave., N.W.
Washington, D.C. 20005
202.371.7000

The White House Secures Voluntary Commitments From Seven Leading AI Companies To Promote Safety, Security and Trust in AI

As the adoption of AI proliferates, a number of different stakeholders have expressed concerns about the lack of transparency in how AI models operate, and the potential harm they might cause if left unchecked. While various pieces of federal legislation regulating AI have been discussed in the United States, these efforts are at their nascent stages. Until such legislation is enacted, industry self-regulation is likely the quickest way to address some of the concerns that have been raised.

To that end, on July 21, 2023, the Biden administration announced that it has secured pledges from — Microsoft, OpenAI, Google, Meta, Amazon, Anthropic and Inflection — to make eight voluntary commitments to promote the safe, secure and trustworthy development and use of AI technology. These voluntary commitments (AI Commitments), discussed below, are available to other companies as well, and are designed to remain in effect until regulations covering substantially the same issues go into effect. Companies may, of course, make additional commitments beyond the eight that were outlined.

The Voluntary AI Commitments

The AI Commitments are grouped into three categories: safety, security and trust. The introductory language to the AI Commitments notes that, when there is a reference to particular models, it applies only to generative models that "are overall more powerful than the current industry frontier" (such as GPT-4 and DALL-E 2).

Safety

1. Companies commit to internal and external red-teaming of models or systems in areas including misuse, societal risks and national security concerns, such as biosecurity, cybersecurity and other safety areas. "Red-teaming" refers to the practice of analyzing a model or system by thinking the way an adversary would and then attempt to attack or challenge a model or system using that perspective.
   - The AI Commitments note that evaluating the safety and capability of AI models, including through the use of red-teaming, is evolving and more work needs to be done. Companies therefore commit to advancing this research, and "developing a multi-faceted, specialized, and detailed red-teaming regime" for all major public releases of new models. This includes research on the interpretability of AI systems' decision-making processes and on increasing the robustness of AI systems against misuse

- Some key areas to focus on in this regard, include:
  - Bio, chemical and radiological risks, including for weapons deployment.
  - Cyber capabilities, including the manner in which AI systems can facilitate attacks by identifying vulnerabilities.
  - Societal risks, such as bias and discrimination.

2. Working toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards.
   - This AI Commitment starts with the premise that information sharing, common standards and best practices are important for advancing AI trust and safety. The companies commit to establishing or joining a forum through which shared standards and best practices for AI safety or future standards related to red-teaming, safety and societal risks can be developed, advanced and adopted.
   - The AI Commitments anticipate companies will be working closely with government, civil society and academia.

Security

3. Investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights.
   - Given the concern about how AI models may be used by others for nefarious purposes, companies commit to treating unreleased AI model weights as core intellectual property with regard to cybersecurity and insider threat risks. The goal here is to get companies to protect these sensitive AI system components the same way they protect their core intellectual property. Examples include storing this information in a secure environment; disclosing this information only to those with a need to know; and establishing an insider threat detection program.
4. Incent third-party discovery and reporting of issues and vulnerabilities.
   - Given that AI systems may have vulnerabilities after release, companies commit to developing incentives to promote the responsible disclosure of weaknesses of AI systems or including AI systems in existing bug bounty programs.

Trust

5. Developing and deploying mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking or both, for AI-generated audio or visual content.
   - A key concern as AI advances is how users will be able to distinguish human-generated and AI-generated works. Companies therefore commit to working with industry peers and standards-setting bodies to develop a framework to help users distinguish between audio-visual content generated by users and audio-visual content generated by AI, such as through the use of watermarking. The watermark should indicate which service or model was used but does not need to include any identifying user information.
   - Companies also commit to developing APIs (application programming interfaces) to determine if a piece of content was generated through their AI system.
6. Publicly reporting model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias.
   - The AI Commitments note that users should understand the capabilities and limitations of AI systems. The companies therefore commit to publishing reports for all new "significant model public releases." The reports are intended to include: the safety evaluations conducted; significant limitations in performance that have implications for appropriate use; discussions of the model's effect on societal risks such as fairness and bias; and results of adversarial testing.
7. Prioritizing research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy.
   - This AI Commitment includes empowering trust and safety teams, advancing AI safety research and privacy, protecting children and striving to proactively manage AI risks.
8. Developing and deploying frontier AI systems to help address society's greatest challenges.
   - This AI Commitment includes supporting initiatives to help citizens understand the nature, capabilities, limitations, and impact of AI technology.

Additional White House AI Initiatives

Securing the AI Commitments from the initial seven participating companies is just the latest step in a broader White House AI initiative. For example, in October 2022, the White House released "The Blueprint for an AI Bill of Rights" (Blueprint), which included five principles and supplementary practices meant to guide the design, use and deployment of automated systems. Although the five principles are largely consistent with the AI Commitments, the Blueprint applies to a broader scope of AI systems. As noted above, the AI Commitments apply only to generative models that "are overall more powerful than the current industry frontier."

However, the Blueprint applies more broadly to "(1) automated systems that (2) have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services."

The White House Secures Voluntary Commitments From Seven Leading AI Companies To Promote Safety, Security and Trust in AI

In addition, in February 2023, President Biden signed an Executive Order that directs federal departments and agencies to eliminate bias in their design and use of new technologies and protect against algorithmic discrimination, and the White House Office of Science and Technology Policy released an updated National AI R&D Strategic Plan in May 2023 that outlines priorities and goals for federal investments in AI research and development.

Next Steps

In announcing the AI Commitments, the White House acknowledged that there is still work to be done and previewed future plans regarding AI. According to the fact sheet that accompanied the AI Commitments, the Biden administration is developing an executive order and plans to pursue bipartisan legislation to promote responsible innovation and protection. Additionally, the administration stated that it will work with allies to develop a strong international code of conduct to govern the development and use of AI globally.

Key Points

Industry self-regulation will be central to how the transparency, safety and security of AI systems are attained in the short to medium term given that it will likely be some time before federal AI legislation is enacted.

As companies navigate this space, it is important to remember the global reach of AI and the development of regulations in other countries. For example, the European Parliament has passed a draft of the EU AI Act, which, broadly speaking, would impose obligations on providers and users of AI systems depending on the level of risk posed by those systems. In the U.K., the government has published a National AI Strategy outlining key priorities with respect to AI, such as investing in the long-term needs of the AI ecosystem, ensuring AI benefits all sectors and regions, and governing AI effectively. In a policy paper published earlier this year, the U.K. government detailed its plans to implement a "pro-innovation approach" to AI regulation.

In the U.S., companies should also be mindful that certain state privacy laws, although not written specifically to address AI, include provisions that may govern how personal data is used and processed with AI systems.

How the AI Commitments are actually implemented remains to be seen, and we expect certain stakeholders in the AI space to argue that the AI Commitments, even if fully implemented, fall short of what is required to manage the risks presented by AI systems.