Ensuring Safe, Secure, and Trustworthy AI

Artificial intelligence offers enormous promise and great risk. To make the most of that promise, America must safeguard our society, our economy, and our national security against potential risks.

The companies developing these pioneering technologies have a profound obligation to behave responsibly and ensure their products are safe.

The voluntary commitments that several companies are making today are an important first step toward living up to that responsibility. These commitments – which the companies are making immediately – underscore three principles that must be fundamental to the future of AI: safety, security, and trust.

> **Safety:** Companies have a duty to make sure their products are safe before introducing them to the public. That means testing the safety and capabilities of their AI systems, subjecting them to external testing, assessing their potential biological, cybersecurity, and societal risks, and making the results of those assessments public.
>
> **Security:** Companies have a duty to build systems that put security first. That means safeguarding their models against cyber and insider threats and sharing best practices and standards to prevent misuse, reduce risks to society, and protect national security.
>
> **Trust:** Companies have a duty to do right by the public and earn the people's trust. That means making it easy for users to tell whether audio and visual content is in its original form or has been altered or generated by AI. It means ensuring that the technology does not promote bias and discrimination, strengthening privacy protections, and shielding children from harm. Finally, it means using AI to help meet society's greatest challenges, from cancer to climate change, and managing AI's risks so that its benefits can be fully realized.

These voluntary commitments are only a first step in developing and enforcing binding obligations to ensure safety, security, and trust. Realizing the promise and minimizing the risk of AI will require new laws, rules, oversight, and enforcement. The Biden-Harris Administration will continue to take executive action and pursue bipartisan legislation to help America lead the way in responsible innovation and protection. As we advance this agenda at home, we will work with allies and partners on a strong international code of conduct to govern the development and use of AI worldwide.

Voluntary AI Commitments

The following is a list of commitments that companies are making to promote the safe, secure, and transparent development and use of AI technology. These voluntary commitments are consistent with existing laws and regulations, and designed to advance a generative AI legal and policy regime. Companies intend these voluntary commitments to remain in effect until regulations covering substantially the same issues come into force. Individual companies may make additional commitments beyond those included here.

Scope: Where commitments mention particular models, they apply only 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).

Safety

1) Commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas.

Companies making this commitment understand that robust red-teaming is essential for building successful products, ensuring public confidence in AI, and guarding against significant national security threats. Model safety and capability evaluations, including red teaming, are an open area of scientific inquiry, and more work remains to be done. Companies commit to advancing this area of research, and to developing a multi-faceted, specialized, and detailed red-teaming regime, including drawing on independent domain experts, for all major public releases of new models within scope. In designing the regime, they will ensure that they give significant attention to the following:

- Bio, chemical, and radiological risks, such as the ways in which systems can lower barriers to entry for weapons development, design, acquisition, or use
- Cyber capabilities, such as the ways in which systems can aid vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system
- The effects of system interaction and tool use, including the capacity to control physical systems
- The capacity for models to make copies of themselves or "self-replicate"
- Societal risks, such as bias and discrimination

To support these efforts, companies making this commitment commit to advancing ongoing research in AI safety, including on the interpretability of AI systems' decision-making processes and on increasing the robustness of AI systems against misuse. Similarly, companies commit to publicly disclosing their red-teaming and safety procedures in their transparency reports (described below).

2) Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards

Companies making this commitment recognize the importance of information sharing, common standards, and best practices for red-teaming and advancing the trust and safety of AI. They commit to

establish or join a forum or mechanism through which they can develop, advance, and adopt shared standards and best practices for frontier AI safety, such as the NIST AI Risk Management Framework or future standards related to red-teaming, safety, and societal risks. The forum or mechanism can facilitate the sharing of information on advances in frontier capabilities and emerging risks and threats, such as attempts to circumvent safeguards, and can facilitate the development of technical working groups on priority areas of concern. In this work, companies will engage closely with governments, including the U.S. government, civil society, and academia, as appropriate.

Security

3) Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights

Companies making this commitment will treat unreleased AI model weights for models in scope as core intellectual property for their business, especially with regards to cybersecurity and insider threat risks. This includes limiting access to model weights to those whose job function requires it and establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets. In addition, it requires storing and working with the weights in an appropriately secure environment to reduce the risk of unsanctioned release.

4) Incent third-party discovery and reporting of issues and vulnerabilities

Companies making this commitment recognize that AI systems may continue to have weaknesses and vulnerabilities even after robust red-teaming. They commit to establishing for systems within scope bounty systems, contests, or prizes to incent the responsible disclosure of weaknesses, such as unsafe behaviors, or to include AI systems in their existing bug bounty programs.

Trust

5) Develop and deploy 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

Companies making this commitment recognize that it is important for people to be able to understand when audio or visual content is AI-generated. To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed. They will also develop tools or APIs to determine if a particular piece of content was created with their system. Audiovisual content that is readily distinguishable from reality or that is designed to be readily recognizable as generated by a company's AI system—such as the default voices of AI assistants—is outside the scope of this commitment. The watermark or provenance data should include an identifier of the service or model that created the content, but it need not include any identifying user information. More generally, companies making this commitment pledge to 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.

6) Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias

Companies making this commitment acknowledge that users should understand the known capabilities and limitations of the AI systems they use or interact with. They commit to publish reports for all new significant model public releases within scope. These reports should include the safety evaluations conducted (including in areas such as dangerous capabilities, to the extent that these are responsible to publicly disclose), significant limitations in performance that have implications for the domains of appropriate use, discussion of the model's effects on societal risks such as fairness and bias, and the results of adversarial testing conducted to evaluate the model's fitness for deployment.

7) Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy

Companies making this commitment recognize the importance of avoiding harmful biases from being propagated by, and discrimination enacted by, AI systems. 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.

8) Develop and deploy frontier AI systems to help address society's greatest challenges

Companies making this commitment agree to 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. Companies also commit to supporting initiatives that foster the education and training of students and workers to prosper from the benefits of AI, and to helping citizens understand the nature, capabilities, limitations, and impact of the technology.