White House Obtains Commitments to Regulation of Generative AI from OpenAI, Amazon, Anthropic, Google, Inflection, Meta and Microsoft

By Ralph Losey, Losey PLLC

Chat Bots say 'Catch me if you can! I move fast.'

In a landmark move towards the regulation of generative AI technologies, the White House brokered eight "commitments" with industry giants Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI. The discussions, held exclusively with these companies, culminated in an agreement on July 21, 2023. Despite the inherent political complexities, all parties concurred on the necessity for ethical oversight in the deployment of their AI products across several broad areas.

Introduction

These commitments, although necessarily ambiguous, represent a significant step to what may later become binding law. The companies not only acknowledged the appropriateness of future regulation across eight distinct categories, they also pledged to uphold their ongoing self-regulation efforts in these areas. This agreement thus serves as a kind of foundation blueprint for future Ai regulation. Also see prior efforts by U.S. government that precede this blueprint, AI Risk Management Framework, (NIST, January 2023), and the White House Blueprint for an AI Bill of Rights, (October 2022).

Figure 1 Government determined to catch and pin down wild chat bots.

The eight "commitments" are outlined in this article with analysis, background and some editorial comments. For a direct look at the agreement, here is a link to the "Commitment" document. For those interested in the broader legislative landscape surrounding AI in the U.S., see my prior article, "Seeds of U.S. Regulation of AI: the Proposed SAFE Innovation Act" (June 7, 2023). It provides a comprehensive overview of proposed legislation, again with analysis and comments. *Also see*, Algorithmic Accountability Act of 2022 (requiring self-assessments of AI tools' risks, intended benefits, privacy practices, and biases) and American Data Privacy and Protection Act (ADPPA) (requiring impact assessments for "large data holders" when using algorithms in a manner that poses a "consequential risk of harm," a category which certainly includes some types of "high-risk" uses of AI).

The document formalizes a voluntary commitment, which is sort of like a non-binding agreement, an agreement to try to reach an agreement. The parties statement begins by acknowledging the potential and risks of artificial intelligence (AI). Then it affirms that companies developing AI should ensure the **safety**, **security**, and **trustworthiness** of their technologies. These are the three major themes for regulation that the White House and the tech companies could agree upon. The document then outlines eight particular commitments to implement these three fundamental principles.

The big tech companies affirm they are already taking steps to ensure the safe, secure, and transparent development and use of AI. So these commitments just confirm what they are already doing. Clever wording here and of course, the devil is always in the details, which will have to be ironed out later as the regulatory process continues. The basic idea that the parties were able to agree upon at this stage is that these eight voluntary commitments, as formalized and described in the document, are to remain in effect until such time as enforceable laws and regulations are enacted.

Figure 2 Just Regulation of Ai Should Be Everyone's Goal.

The scope of the eight commitments is specifically limited to generative Ai models that are more powerful than the current industry standards, specified in the document as, or equivalent to: GPT-4, Claude 2, PaLM 2, Titan, and DALL-E 2 for image generation. Only these models, or models more advanced than these, are intended to be covered by this first voluntary agreement. It is likely that other companies will sign up later and make these same general commitments, if nothing else, to claim that their generative technologies are now of the same level as these first seven companies.

It is a good for discussions like this to start off in a friendly manner and reach general principles of agreement on the easy issues – the low hanging fruit. Everyone wants Ai to be **safe**, **secure**, and **trustworthy.** The commitments lay a foundation for later, much more challenging discussions between industry and government and the people the government is supposed to represent. Good work by both sides in what must have been very interesting opening talks.

Figure 3 What can we agree upon to start talking about regulation?

Dissent in Big Tech Ranks Already?

It is interesting to see that there is already a split among the seven big tech companies whom the White Hours talked into the commitments, Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI. Five of them went on to create an industry group focused on ensuring safe and responsible development of frontier AI models, which they call the Frontier Model Forum (announced July 26, 2023). Two did not join the Forum: Amazon and Inflection. And you cannot help but wonder about Apple, who apparently was not even invited to the party at the White House, or maybe they were, and decided not to attend. Apple should be in these discussions, especially since they are rumored to be well along in preparing a advanced Ai product. Apple is testing an AI chatbot but has no idea what to do with it, (Verge, July 19, 2023).

Inflection AI, Inc., the least known of the group, is a $4 billion private start-up that claims to have the world's best AI hardware setup. *Inflection AI, The Year-Old Startup Behind Chatbot Pi, Raises $1.3 Billion*, (Forbes, 6/29/23). Inflection is company behind the empathetic software, PI, which I previously wrote about in Code of Ethics for "Empathetic" Generative AI, (July 12, 2023). These kind of personal, *be your best friend*, chat bots present special dangers of misuse, somewhat different than the rest. My article delves into this and endorses Jon Neiditz' proposed *Code of Ethics for "Empathetic" Generative AI*.

Control Promotion and Exploitation of Robot Love.

The failure of Inflection to join in the Frontier Model Forum is concerning. So too is Amazon's recalcitrance, especially considering the number of Alexa ears there are in households world wide (I have two), not to mention their knowledge of most everything we buy.

Think Universal, Act Global

The White House Press Release on the commitments says the Biden Administration plans to "*continue executive action and pursue bipartisan legislation for responsible innovation and protection.*" The plan is to, at the same time, work with international allies to develop a code of conduct for AI development and use worldwide. This is ambitious, but appropriate for the U.S. government to think globally on these issues.

The E.U. is already moving fast in Ai regulation, many say too fast. The E.U. has a history of strong government involvement with big tech regulation, again, some say too strong, especially on the E.U.'s hot button issue, consumer privacy. *The EU and U.S. Diverge on AI Regulation: A Transatlantic Comparison and Steps to Alignment*, (Brookings Institution, 2/16/23). I am inclined towards the views of privacy expert, Jon Neiditz, who explains why generative Ais provide significantly **more** privacy than the existing systems. *How to Create Real Privacy & Data Protection with LLMs*, (The Hybrid Intelligencer, 7/28/23) ("… *replacing Big Data technologies with LLMs can create attractive, privacy enhancing alternatives to the surveillance with which we have been living.*") Still, privacy in general remains a significant concern for all technologies, including generative Ai.

The free world must also consider the reality of the technically advanced totalitarian states, like China and Russia, and the importance to them of Ai. *Artificial Intelligence and Great Power Competition, With Paul Scharre*, (Council on Foreign Relations ("CFR"), 3/28/23) (Vladimir Putin said in September 2017: "*Artificial intelligence is the future not only for Russia, but for all humankind. Whoever becomes the leader in this sphere will become the ruler of the world.*" . . . *[H]alf of the world's 1 billion surveillance cameras are in China, and they're increasingly using AI tools to empower the surveillance network that China's building*); *AI Meets World, Part Two*, (CFR, June 21, 2023) (good background discussion on Ai regulation issues, although some of the commentary and questions in the audio interview seem a bit biased and naive).

There is a military and power control race going on. This makes U.S. and other free-world government regulation difficult and demands *eyes wide open* international participation. Many analysts now speak of the need for global agreements along the lines of Nuclear Non-Proliferation treaties attained in the past. *See eg*., *It is time to negotiate global treaties on artificial intelligence*, (Brookings Institute, 3/24/21); *OpenAI CEO suggests international agency like UN's nuclear watchdog could oversee AI*, (AP, 6/6/23); *But see*, *Panic about overhyped AI risk could lead to the wrong kind of regulation*, (Verge, 7/3/23).

Mad Would Be World Dictators Covet Ai.

Three Classes of Risk Addressed in the Commitments

Safety. Companies are all expected to ensure their AI products are safe before they are introduced to the public. This involves testing AI systems for their safety and capabilities, assessing potential biological, cybersecurity, and societal risks, and making the results of these assessments public. *See*: *Statement on AI Risk*, (Center for AI Safety, 5/30/23) (open letter signed by many Ai leaders, including Altman, Kurzweil and even Bill Gates, agreeing to this short statement "*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.*"). The Center for AI Safety provides this short statement of the kind of societal-scale risks it is worried about:

> AI's application in warfare can be extremely harmful, with machine learning enhancing aerial combat and AI-powered drug discovery tools potentially being used for developing chemical weapons. CAIS is also concerned about other risks, including increased inequality due to AI-related power imbalances, the spread of misinformation, and power-seeking behavior.
>
> *FAQ of Center for AI Safety*

These are all very valid concerns. The spread of misinformation has been underway for many years.

The disclosure requirement will be challenging in view of both competitive and intellectual property concerns. There are related criminal hacking and military concerns that disclosure and open source code may help criminal hackers and military espionage. Michael Kan, *FBI: Hackers Are Having a Field Day With Open-Source AI Programs* (PC Mag., 7/28/23) (Criminals are using AI programs for phishing schemes and to help them create malware, according to a senior FBI official). Foreign militaries, such as China and Russia are known to be focusing on Ai technologies for suppression and attacks.

The commitments document emphasizes the importance of external testing and the need for companies to be transparent about the safety of their AI systems. The external testing is a good idea and hopefully this will be by an independent group, and not just the leaky government, but again, there is the transparency concern with over-exposure of secrets and China's well-known constant surveillance and theft of IP.

Figure 4 Testing new advanced Ai products before release to public

Note the word "license" was not used in the commitments, as that seems to be a *hot button* for some. *See eg*. *The right way to regulate AI*, (Case Text, July 23, 2023) (claims that Sam Altman proposed no one be permitted to work with AI without first obtaining a license). With respect, that is not a fair interpretation of Sam Altman's Senate testimony or OpenAI's position. Altman talked said "*licensing and testing of all Ai models.*" This means **licensing of Ai models** to confirm to the public that the models have been tested and approved as safe. In context, and based on Altman's many later explanations in his world tour that followed, it is obvious that Sam Altman, OpenAI's CEO, meant a license to sell a particular product, not a license for a person to work with Ai at all, nor a license to create new products, or do research. *See eg*. the lengthy video interview of Sam Altman given to Bloomberg Technology on June 22, 2026.

Regulatory *licensing* under discussion so far pertains only to the final products, to certify to all potential users of the new Ai tech that it has been tested and certified as safe, secure, and trustworthy. Also the license scope would be limited to very advanced new products, which do, almost all agree, present very real risks and dangers. No one wants a new FDA, and certainly no one wants to require individual licenses for someone to use Ai, like a driver's license, but it seems like common sense to have these powerful new technology products tested and approved by some regulatory body before a company releases it. Again, the devil in in the details and this will be a very tough issue.

Figure 5 Keeping Us Safe.

Security. The agreement highlights the duty of companies to prioritize security in their AI systems. This includes safeguarding their models against cyber threats and insider threats. Companies are also encouraged to share best practices and standards to prevent misuse of AI technologies, reduce risks to society, and protect national security. One of the underlying concerns here is how Ai can be used by criminal hackers and enemy states to defeat existing *blue team* protective systems. Plus, there is the related threat of commercially driven races of Ai products to the market before they are ready. Ai products need adequate *red team* testing before release, coupled with ongoing testing after release. The situation is even worse with third-party plug-ins. They often have amateurish software designs and no real security at all. In today's world, cybersecurity must be a priority of everyone. More on this later in the article.

Figure 6 AI Cyber Security.

Trust. Trust is identified as a crucial aspect of AI development. Companies are urged to earn public trust by ensuring transparency in AI-generated content, preventing bias and discrimination, and strengthening privacy protections. The agreement also emphasizes the importance of using AI to address societal challenges, such as cancer and climate change, and managing AI's risks so that its benefits can be fully realized. As frequently said on the e-Discovery Team blog, "trust but verify." That is where testing and product licensing come in. For instance, how else would you really know that any confidential information you use with an Ai product is in fact kept confidential as the seller claims? Users are not in a position to verify that. Still, generative Ai is an inherently more privacy protective tech system than existing Big Data surveillance systems. *How to Create Real Privacy & Data Protection with LLMs*.

Figure 7 Ready to Trust Generative Ai?

Eight Commitments in the Three Classes

First, here is the quick summary of the eight commitments:

1. Internal and external *red-teaming* of models,
2. Sharing information about trust and safety risks,
3. Investing in cybersecurity,
4. Incentivizing third-party discovery of vulnerabilities,
5. Developing mechanisms for users to understand if content is AI-generated,
6. Publicly reporting model capabilities and limitations,
7. Prioritizing research on societal risks posed by AI,
8. Deploying AI systems to address societal challenges.

Figure 8 Preparing Early Plans for Ai Regulation.

Here are the document details of the eight commitments, divided into the three classes of risk. A few e-Discovery Team editorial comments are also included and, for clarity, are shown in **(bold parenthesis**).

Two Safety Commitments

1. Companies commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns. **(This is the basis for the President Biden's call for hackers to attend DEFCON 31 to "red team" and expose errors and vulnerabilities that experts in Ai discover in open competitions. We will be at DEFCON to cover these events. *Vegas Baby! DEFCON 31*.)** The companies all acknowledge that robust red-teaming is essential for building successful products, ensuring public confidence in AI, and guarding against significant national security threats. **(An example of new employment opportunities made possible by Ai.)** The companies also commit to advancing ongoing research in AI safety, including the interpretability of AI systems' decision-making processes and increasing the robustness of AI systems against misuse. **(Such research is another example of new work creation by Ai.)**

2. Companies commit to work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards. **(Such information sharing is another example of new work creation by Ai.)** They 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. **(Another example of new, information sharing work created by Ai. These forums all require dedicated human administrators.)**

Figure 9 Everyone Wants Ai to be Safe.

Two Security Commitments

3. On the security front, companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights. The companies treat unreleased AI model weights as core intellectual property, 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. **(Again, although companies already invest in these jobs, even more work, more jobs, will be created by these new AI IP related security challenges, which will, in our view, be substantial. We do not want enemy states to steal these powerful new technologies. The current cybersecurity threats from China, for instance, are already extremely dangerous, and may encourage their attack of Taiwan, a close ally who supplies over 90% of the world's advanced computer chips. *Taiwan's dominance of the chip industry makes it more important*, (The Economist, 3/16/23); *U.S. Hunts Chinese Malware That Could Disrupt American Military Operations*, (NYT, 7/29/23))**.

4. Companies also commit to incentivizing third-party discovery and reporting of issues and vulnerabilities, recognizing that AI systems may continue to have weaknesses and vulnerabilities even after robust red-teaming. **(Again, this is the ongoing Red Teaming mentioned to incentivize researchers, hackers all, to find and report mistakes in Ai code. There have been a host of papers and announcements on Ai vulnerabilities and red team successes lately. *See eg*.: Zou, Wang, Kolte, Fredrikson, Universal and Transferable Attacks on Aligned Language Models, (July 27, 2023); Pierluigi Paganini, *FraudGPT, a new malicious generative AI tool appears in the threat landscape*, (July 26, 2023) (dangerous tools already on dark web for criminal hacking). Researchers should be paid rewards for this otherwise unpaid work. The current rewards should be increased in size to encourage the often not fully employed, economically disadvantaged hackers to do the right thing. Hackers who find errors and succumb to temptation and use them for criminal activities should be punished. There are always errors in new technology like this. There are also a vast number of additional errors and vulnerabilities created by third-party plugins in the gold rush to Ai profiteering. *See eg*: *Testing a Red Team's Claim of a Successful "Injection Attack" of ChatGPT-4 Using a New ChatGPT Plugin*, (May 22, 2023). Many of the mistakes are already well known and some are still not corrected. This appears like inexcusable neglect and we expect future hard laws to dig into this much more deeply. All companies need to be ethically responsible and the big Ai companies need to police the small plug-in companies, much like Apple now polices its App Store. We think this area is of critical importance.)**

Figure 10 Guard Against Ai "Prison Breaks"

Four Trust Commitments

5. In terms of trust, companies commit to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated. This includes developing strong mechanisms, such as provenance and/or **watermarking systems** for audio or visual content created by any of their publicly available systems. **(This is a tough one, and only will grow in importance and difficulty as these systems grow more sophisticated. OpenAI experimented with watermarking, but were disappointed at the results and quickly discontinued it. *OpenAI Retires AI Classifier Tool Due to Low Accuracy*, (Fagen Wasanni Technologies, July 26, 2023). How do we even know if we are actually talking to a person, and not just an Ai posing as a human? Sam Altman has launched a project outside of OpenAI addressing that challenge, among other things, the World Coin project. On July 27, 2023, they began to verify that an online applicant to World Coin membership is in fact human. They do that with in-person eye scans in physical centers around the world. An interesting example of new jobs being created to try to meet the 'real or fake' commitment.)**

6. Companies also commit to publicly reporting model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of the model's effects on

societal risks such as fairness and bias. **(Again, more jobs and skilled human workers will be needed to do this.)**

7. Companies prioritize research on societal risks posed by AI systems, including avoidance of harmful bias and discrimination, and protection of privacy. **(Again, more work and employment. Some companies might prefer to gloss over and minimize this work because it will slow and negatively impact sales, at least at first. Glad to see these human rights goals in an initial commitment list. We expect the government will set up extensive, detailed regulations in this area. It has a strong political, pro-consumer draw.)**

8. Finally, companies commit to developing and deploying frontier AI systems to help address society's greatest challenges. These challenges include climate change mitigation and adaptation, early cancer detection and prevention, and combating cyber threats. They 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. **(We are big proponents of this and the possible future benefits of Ai. See eg, ChatGTP-4 Prompted To Talk With Itself About "The Singularity", (April 4, 2023), and Sam Altman's Favorite Unasked Question: What Will We Do in the Future After AI?, (July 7, 2023))**.

Figure 11 Totally Fake Image of Congressman Lieu (pretty obvious to most, even without watermarks).

Conclusion

The Commitments document emphasizes the need for companies to take responsibility for the safety, security, and trustworthiness of their AI technologies. It outlines eight voluntary commitments to advance the principles. The voluntary agreement highlights the need for ongoing research, transparency, and public engagement in the development and use of AI. The e-Discovery Team blog is already doing its part on the "public engagement" activity, as this is our 38th article in 2023 on generative Ai.

The Commitments document closes by noting the potential of AI to address some of society's greatest challenges, while also acknowledging the risks and challenges that need to be managed. It is important to do that, to remember we must strike a fair balance between protection and innovation. Seeds of U.S. Regulation of AI: the Proposed SAFE Innovation Act.

The e-Discovery Team blog always tries to do that, in an objective manner, not tied to any one company or software product. Although ChatGPT-4 has so far been our clear favorite, and their software is the one we most frequently use and review, that can change, as other products enter the market and improve. We have no economic incentives or secret gifts tipping the scale of our judgments.

Although some criticize the Commitments as meaningless showmanship, we disagree. From Ralph's perspective as a senior lawyer, with a lifetime of experience in legal negotiations, it looks like a good start and show of good faith on both sides, government and corporate. We all want to control and prevent Terminator robot dystopias.

Figure 12 Figure 12 Justice depends on reasoning free from a judge's personal gain

Figure 13 Lawyer stands over Terminator robot he just defeated.

Still, it is just a start, far from the end goal. We have a long way to go and naive idealism is inappropriate. We must *trust and verif*y. We must operate in the emerging world with eyes wide open. There are always conmen and power-seekers seeking to profit from new technologies. Many are motivated by what Putin said about Ai: "*Whoever becomes the leader in this sphere will become the ruler of the world.*"

Figure 14 Trust But Verify!

Many believe AI is, or may soon be, the biggest technological advance of our age, perhaps of all time. Many say it will be bigger than the internet, perhaps equal to the discovery of nuclear energy. Just as Einstein's discovery, with Oppenheimer's engineering, resulted in the creation of nuclear weapons that ended WWII, these discoveries also left us with an endangered world living on the brink of total thermonuclear war. Although we are not there yet, Ai creations could eventually take us to the same DEFCON threat level. We need Ai regulation to prevent that.

Governments word-wide must come to understand that using Ai as an all out, uncontrolled weapon will result in a war game that cannot be won. It is a Mutually Assured Destruction ("MAD") tactic. The global treaties and international agencies on nuclear weapons and arms control, including the military use of viruses, were made possible by the near universal realization that nuclear war and virus weapons were MAD ideas.

Figure 15 MAD AI War Apocalypse

All governments must be made to understand that everyone will lose an Ai world war, even the first strike attacker. These treaties and inspection agencies and MAD realization have, so far enabled us to avoid such wars. We must do the same with Ai. Governments must be made to understand the reality of Ai triggered species extermination scenarios. Ai must ultimately be regulated, bottled up, on an international basis, just as nuclear weapons and bioweapons have been.

Ai must be regulated to prevent uncontrollable consequences.