Written testimony of

Helen Toner
Director of Strategy and Foundational Research Grants
Center for Security and Emerging Technology, Georgetown University

Before the U.S. Senate Committee on the Judiciary
Subcommittee on Privacy, Technology, and the Law

For a hearing on "Oversight of AI: Insiders' Perspectives"
September 17, 2024

—

Chair Blumenthal, Ranking Member Hawley, members of the subcommittee: thank you for the opportunity to testify today. I commend you for the depth with which you are approaching AI policy issues and the sustained focus you have had on the varied challenges and opportunities this technology brings. My testimony will focus on one particular set of topics where, in my view, the gap between the discourse inside leading AI companies and the discourse in policy circles is largest.

My background is as follows: I have spent the last 8 years working on AI policy, focusing on national security issues, US-China competition, and AI safety. During that time I lived in San Francisco, then Beijing, before moving to Washington in 2019 to help found the Center for Security and Emerging Technology (CSET) at Georgetown University, which has since grown to be one of policymakers' most respected sources of analysis on national security and AI. While working at CSET, I also served on the nonprofit board that governs OpenAI, from 2021 until the well-documented events of last November. In my testimony today, I seek to offer a perspective on AI policy and regulatory issues that is informed by my 5 years in the leadership team at CSET, my 2.5 years on the OpenAI board, and the extensive relationships I have developed with

researchers and leaders within the AI industry over the years I have worked in this space. The views I express are my own, and do not represent any organization with which I am affiliated.

The biggest disconnect I see between public perceptions and AI insider perspectives comes from inside the handful of companies that are working to build "artificial general intelligence" (AGI), i.e. AI that is roughly as smart as a human. Companies with this explicit goal include OpenAI, Google, Anthropic, Meta, Amazon, Microsoft, and xAI. **In public and policy conversations, talk of AGI is often treated as either a science-fiction pipe dream or a marketing ploy. Among the scientists and engineers of these companies, it is an entirely serious goal**—one that many involved see as achievable within 10 or 20 years, and some in as little as 1-3 years. Even if the shortest estimates turn out to be wrong, **the idea of human-level AI being developed in the next decade or two should be seen as a real possibility that necessitates significant preparatory action now.**

Policy conversations about AGI often get derailed by definitional issues: the concept of "AGI" is highly contested, with different experts offering very different assessments of what would count as AGI and how feasible it is to build such a system. The same applies to "superintelligence"—loosely, an AI system that is far smarter or more capable than any human.[1] But for the purposes of making policy today, nailing down the perfect definition is not what matters.

What matters is that billions of dollars and thousands of the world's brightest minds are being directed towards building machines that can perform an increasing proportion of tasks that we used to think could only be done by humans, and that these efforts seem to be making progress. Such vast resources are being committed because many top researchers and engineers believe that the last 10+ years of AI progress show a steady trend towards more capable AI systems, with plenty of room to get far more capable still. From the outside, it may have looked like ChatGPT represented a one-off leap in what AI can do. To the researchers who built it, ChatGPT was just one stop along a road that they have been on for years and that they expect to continue bearing fruit.

The problem is, the same people expecting to build advanced AI in the coming years believe that doing so could be incredibly dangerous. Experts agree that at a minimum, if we build AI systems that are smarter than humans in ways that matter, this technology will radically transform society. At a minimum, it will be an enormously powerful

[1] *"There is no fundamental reason for AI progress to slow or halt at human-level abilities. […] Compared with humans, AI systems can act faster, absorb more knowledge, and communicate at a higher bandwidth. Additionally, they can be scaled to use immense computational resources and can be replicated by the millions. We do not know for certain how the future of AI will unfold. However, we must take seriously the possibility that highly powerful generalist AI systems that outperform human abilities across many critical domains will be developed within this decade or the next. What happens then?"* Bengio et al. 2024, "Managing extreme AI risks amid rapid progress."

tool that could do huge harm in the wrong hands. Many CEOs and respected experts go further: they believe that if we build AI that is smarter than us, it is possible—even likely—that we will be unable to control it, which will result in disaster.[2] Views differ on whether "disaster" means literal human extinction or something more pernicious, like humans slowly handing over more and more control to machines until the morning when we wake up and realize we have no control over our own future. Either way, the point is clear: **this technology would be enormously consequential, potentially extremely dangerous, and should only be developed with careful forethought and oversight.** It is jarring to see the mismatch between 1) internal expectations of how consequential these companies think the technology they are building will be, and 2) how seriously they are approaching questions about how safe their AI systems need to be and who gets to decide how they are developed and used.

The following facts about advanced AI development today, taken together, should be quite concerning to this committee and to policymakers generally:

- **The science of measuring AI progress and AI risks is extremely immature.** We do not have scientifically grounded, agreed-upon methods to determine how rapidly AI development is progressing towards AGI, how to compare the capabilities (or "intelligence" levels) of different AI models, or how to assess what risks they pose. While it is clear that today's AI systems are far more capable than they were 5 or 10 years ago, experts disagree wildly on what to expect in the future. There does not exist a high-quality, neutral methodology to adjudicate between views.
- **The science of managing risks and harms from AI is similarly immature.** AI companies use a range of ad-hoc approaches to handle problems as they arise. Current safety methods—such as using reinforcement learning to make chatbots less likely to say things they aren't supposed to say—are bandaids, not fundamental fixes, and are unlikely to hold up as the underlying systems get smarter. Some researchers are pursuing potentially promising avenues to develop better techniques, but this work will take time to bear fruit, and is currently far less resourced than efforts to make AI smarter.
- **Companies working towards advanced AI are subject to enormous pressure to move fast, beat their competitors to market, and raise money from investors.** The current paradigm of AI development demands huge amounts of capital—billions of dollars, maybe tens or hundreds of billions—to build and run data centers containing hundreds of thousands of cutting-edge chips. There is also fierce competition over top talent, which companies can only attract if they are seen as the hottest place to do cutting-edge work.

---

[2] Dario Amodei, Anthropic CEO: *"My chance that something goes, you know, really quite catastrophically wrong on the scale of human civilization might be somewhere between 10-25%."*
Sam Altman, OpenAI CEO: *"The bad case [...] is, like, lights out for all of us."*
Geoffrey Hinton, Turing Award winner: *"I think we've got a better than even chance of surviving it. But it's not like there's only a 1% chance of [superintelligence] taking over. It's much more than that."*
Ilya Sutskever, OpenAI co-founder: *"The future is going to be good for the AIs regardless. It would be nice if it would be good for humans as well."*

The result of these factors is that even the companies that are trying hard to do the right thing *do* have a strong incentive to move as fast as they can, and *don't* have any clarity around what kinds of safety testing and mitigation are sufficient for their latest and greatest release. It's unsurprising that in some companies, this leads to **product launch dates dictating how much safety work can be crammed in beforehand, rather than being dictated by when a clear safety bar is reached.** Imagine if rocket launches were planned the same way. My experience on the board of OpenAI taught me how fragile internal guardrails are when money is on the line, and why it's imperative that policymakers step in.

AI companies often argue that it is too early to implement any kind of AI regulation, because the science of making these systems safe is still nascent. But science or no science, those same companies keep finding ways to turn the crank on money and talent in order to create increasingly capable systems.[3] A "wait and see" approach to policy is totally inadequate given that **these systems are being built and deployed—and affecting hundreds of millions of people's lives—even in the absence of scientific consensus about how they work or what will be built next.** Smart, flexible, and foresighted AI policy can start to manage risks and harms now, while also putting us in a better position to respond to changes in the field over time.

I want to be clear: I do not know how long we have to prepare for smarter-than-human AI, and I don't know how hard it will turn out to be to control it and ensure it is safe. As I'm sure the committee has heard a thousand times at this point, AI doesn't just bring risks—it also has the potential to raise living standards, help solve global challenges, and empower people around the world. **If the story were simply that this technology is bad and dangerous, policymakers' task would be easy.** The challenge we face is figuring out how to navigate immense uncertainty about how quickly AI will progress, what dangers will arise along the way, and how long society will have to adapt to disruptions and changes.

The good news is that there are **light-touch, adaptive policy measures** that can not only help navigate the issues I've focused on here—they would also help tackle many of the present-day challenges we are already experiencing from less advanced AI systems. These policies include:

- **Implement basic building blocks** that can help with a wide range of different AI issues:[4]

[3] *"LLMs predictably get more capable with increasing investment, even without targeted innovation… Many important LLM behaviors emerge unpredictably as a byproduct of increasing investment… There are no reliable techniques for steering the behavior of LLMs… Experts are not yet able to interpret the inner workings of LLMs… Human performance on a task isn't an upper bound on LLM performance."* Bowman 2023, "Eight Things to Know about Large Language Models."

[4] For more on these policies, see Arnold and Toner 2024.

    - **Set transparency requirements** for developers of high-stakes AI systems, including regarding training data, capability testing, safety testing, risk management practices, internal deployments, safety cases, and real-world incidents. Requirements of this kind would significantly shrink the information gap between top companies and government, thereby putting lawmakers in a significantly better position to notice and respond to future progress in the technology—a crucial prerequisite for adaptive, forward-looking policy.
        - One particularly concrete measure would be to codify nascent reporting requirements for advanced AI systems into statute (see recent BIS request for comment).
    - Fund and promote the development of **AI measurement science and safety research** via NIST, NSF, DOE, and other agencies. Our limited ability to understand and control deep learning-based systems lies at the root of many of AI's risks and harms, but these systems are not magical. Sustained support for fundamental research into how deep learning works and how we can measure progress could make a huge difference in our ability to govern this technology.
    - Actively support the development of a **rigorous 3rd-party audit ecosystem**, for example by requiring audits for some AI systems and establishing a federal authority that can license auditors. Independent audits and certifications are central to effective regulation in many sectors; in their absence, lawmakers are forced to rely on companies' own representations of what they are doing and why it is safe. Independent auditors also offer a middle ground between total secrecy and total transparency: companies can share confidential information with selected, vetted partners in order to validate that they are being truthful and responsible.
    - Bolster **whistleblower protections** for employees of AI companies, to ensure that they have clear legal channels to raise safety concerns that are not covered by existing whistleblower law.
    - Increase **technical expertise in government** by streamlining federal hiring processes and adequately resourcing agencies with technical roles to play (e.g. NIST).
    - Clarify how **liability for AI harms** should be allocated, to ensure AI developers and deployers are incentivized to take reasonable care when their products carry a risk of causing serious damage.
- **Lay the groundwork to be able to ramp up government oversight as AI systems become more advanced** over the coming years. A major goal of the above-described policies should be to enable the federal government to notice and respond as AI gets more sophisticated. Congress should be working to develop more robust oversight mechanisms that can be implemented over time—perhaps quickly,

perhaps slowly—as needed.[5] In this vein, I commend Senators Blumenthal and Hawley on their work to build out their bipartisan framework for AI legislation.

A closing note on China: The specter of ceding U.S. technological leadership to China is often treated as a knock-down argument against implementing regulations of any kind. Based on my research on the Chinese AI ecosystem and U.S.-China technology competition more broadly, I think this argument is not nearly as strong as it seems at first glance. We should certainly be mindful of how regulation can affect the pace of innovation at home, and keep a close eye on how our competitors and adversaries are developing and using AI. But looking in depth at Chinese AI development, the AI regulations they are already imposing, and the macro headwinds they face leaves me with the conclusion that they are far from being poised to overtake the United States.[6] The fact that targeted, adaptive regulation does not have to slow down U.S. innovation—and in fact can actively support it—only strengthens this point.

Thank you, and I look forward to your questions.

[5] "If-then commitments" are a helpful emerging framework for this kind of oversight. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework are voluntary, industry-led examples of if-then commitments.
[6] For more, see Toner, Xiao, and Ding 2023.