To: National AI Advisory Committee
From: Thomas Larsen, Executive Director, Center for AI Policy
Date: 10/18/2023

Thank you for the opportunity to share my perspective and concerns regarding the rapid evolution of AI and its implications for our national security. It is an honor to engage with members of this committee, and I am hopeful that we can navigate the challenges ahead by developing a deep understanding of the risks and taking appropriate action in the coming months.

The Center for AI Policy is an advocacy and research organization dedicated to reducing the catastrophic risks from future advanced AI systems. Our team has experience in AI safety research, communications, policy, and law. We believe that there is a significant chance that in the next 3–10 years, AI systems will pose significant threats to national security. We're developing and advocating for policies that would prevent such dangerous AI systems from being created and deployed. Our current policy recommendations include compute governance, frontier AI regulation, and international cooperation on AI.

Catastrophic AI capabilities could emerge this decade

In the past 10 years, one of the primary drivers of AI capabilities has been scaling up algorithms with more computational resources. During this time, the amount of compute used to train advanced AI systems has increased at a rate of about 4.2x per year. Current large training runs cost $100M, and some AI projects are already planning $1B training runs in 2024. Several top AI companies, receiving billions of dollars of investment, are aiming to build smarter-than-human AI. AI systems will soon pass human-level performance on benchmarks in mathematics, logical reasoning, and computer programming.

Additionally, AI is already being used to automate portions of AI research. In the coming years, new AI breakthroughs will produce stronger AI systems that can contribute even more effectively to AI research. Furthermore, improved AI products will draw increased financial investment into AI research. Due to these self-reinforcing trends, economic growth models forecast that within the next decade there may be massive leaps in AI capabilities unfolding over the span of mere months.

> *The idea that this stuff could actually get smarter than people [...] I thought it was way off. I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that.*
> — Geoffrey Hinton, Turing Award winner

The timeline to smarter-than-human AI systems that could pose catastrophic threats is a subject of ongoing debate in the AI community. It is difficult to make precise forecasts about future capabilities, and past predictions have been wrong with both underestimates and overestimates. However, given current progress and development dynamics, it is unwise to be highly confident that smarter-than-human AI is far away.

Misalignment and weaponization risk

Smarter-than-human AI systems would pose serious national security risks. A key subset of the risks come from misaligned AI: AI systems that advance objectives contrary to their designers' intentions. There are a number of threat models for how this might come about, including "rogue" AI systems that act autonomously and deliberately to cause large-scale harm. For a good overview of catastrophic risks from misaligned AI systems, see Google DeepMind's "Threat Model Literature Review." Many AI experts believe that alignment of smarter-than-human systems is an unsolved problem:

> *We do not know, and probably aren't even close to knowing, how to align a superintelligence.*
> — Sam Altman, OpenAI CEO

Aside from misalignment concerns, bad actors could weaponize AI. For example, actors could use future AI systems to develop biological weapons, increase their military capabilities, or conduct large-scale cyberattacks on critical infrastructure. Early indicators of dangerous capabilities are already starting to appear in the development of biological weapons, chemical weapons, and cyber weapons. AI capabilities will only increase from here, and future AI systems may begin significantly assisting with creating weapons of mass destruction.

Policies to mitigate these threats

Addressing catastrophic threats from AI should be a national security priority of the United States. The government needs to prevent the development and deployment of dangerous AI models, both domestically and abroad. Meanwhile, the U.S. should encourage innovation in AI safety to allow for the development of AI systems that can provide massive benefits, while avoiding catastrophic risks.

My team at the Center for AI Policy supports the following proposals: implementing **controls on hardware**, **regulating frontier AI development**, and promoting **international cooperation on AI**.

Controls on hardware: Hardware controls are key because computational power is the difficult-to-obtain ingredient that has driven the current wave of AI. With access to sufficient AI hardware, entities can build powerful AI systems; without this access, it is impractical to build powerful AI systems. An initial step towards robust hardware controls is building a registry of advanced AI chips that tracks basic information like location, ownership, use case, and cluster size.

Domestic regulatory regime: The government needs the capacity to rapidly identify and respond to AI threats. This is why we're advocating for the establishment of a federal office focused on frontier AI risks. Companies should need to apply for a license from the regulator before they can develop or deploy a new frontier AI model. Licenses should be granted based on whether the companies' planned safety and security measures are adequate for the anticipated risks, both from weaponization and misalignment. Beyond licensing, the regulator should monitor AI development and risks. If it identifies a clear national security emergency, it should be empowered to pause dangerous AI projects until safeguards can be put into place.

Some AI companies have existing voluntary standards, such as Anthropic's Responsible Scaling Policy. Unfortunately, voluntary standards will not be adopted by everyone and may be too weak to prevent catastrophes, in large part due to the financial incentive to continue scaling even when AI systems pose severe risks, so the government should step in to require sufficient safeguards.

International cooperation: The U.S. should aim for international cooperation with nations agreeing to prevent the development of smarter-than-human AI systems until adequate safety techniques are developed. Cooperation is critical to avoid a "race to the bottom" in AI, where each actor invests less into safety in order to gain a competitive edge. For non-cooperative countries, the U.S. should use hardware export controls to cripple their ability to build advanced AI systems by preventing them from importing or building production capacity for advanced chips. To enforce these controls, the Bureau of Industry and Security (BIS) at the Department of Commerce should receive increased funding.

Conclusion

Thank you for considering the proposals outlined in this document. The rapidly advancing landscape of AI presents both challenges and opportunities. With the right policies in place, we can mitigate risks of this transformative technology, allowing us to reap the benefits. I look forward to engaging further with this esteemed committee as it safeguards our national interests.