# Written Testimony of Dario Amodei, Ph.D.
# Co-Founder and CEO, Anthropic

## For a hearing on "Oversight of A.I.: Principles for Regulation"

## Before the Judiciary Committee
## Subcommittee on Privacy, Technology, and the Law
## United States Senate
## July 25th, 2023

**Introduction**

Chairman Blumenthal, Ranking Member Hawley, and Members of the Committee, thank you for the opportunity to discuss the risks and oversight of AI with you. I'm Dario Amodei, CEO of Anthropic. Anthropic is a public benefit corporation that aims to lead by example in developing and publishing techniques to make AI systems safer and more controllable, and deploying those techniques thoughtfully in state of the art models.

Research conducted by Anthropic includes *constitutional AI*, a method for training an AI system to behave according to a set of explicit principles; early work on *red teaming*, or adversarial testing of AI systems to uncover bad behavior, a concept which has played a prominent role in the voluntary commitments announced by seven leading AI companies Friday; and a series of foundational works in *AI interpretability*, the science of trying to understand why AI systems behave the way they do.

This month, after extensive testing, we were proud to launch our AI model Claude 2 for U.S. users. Claude 2 puts many of these safety innovations into practice. While we're the first to admit that our measures are still far from perfect, we believe they are an important contribution towards a "race to the top" on safety. We hope we can inspire others in the industry to raise the bar even further.

I will devote most of this prepared testimony to discussing the risks of AI, including what I believe to be extraordinarily grave threats to US national security over the next 2 to 3 years. But before I do that, I wanted to answer one obvious question up front: if I truly believe that AI's risks are so severe, why even develop the technology at all?

To this I have three answers: first, if we can mitigate the risks of AI, its benefits will be truly profound. In the next few years it could greatly accelerate treatments for diseases such as cancer, lower the cost of energy, revolutionize education, improve efficiency throughout government, and much more. Second, relinquishing this technology in the United States would simply hand over its power, risks, and moral dilemmas to adversaries who do not share our values. Finally, a consistent theme of our research has been that the best mitigations to the

risks of powerful AI often *also* involve powerful AI. In other words, the danger and the solution to the danger are often coupled. Being at the frontier thus puts us in a strong position to develop safety techniques (like those I've mentioned above), and also to see ahead and warn about risks, as I'm doing today.

**The Pace of AI Progress**

The single most important thing to understand about AI is how fast it is moving. I have personally never seen anything resembling this pace of progress, and many scientists with longer careers than I seem to concur. Further, the progress is *predictable* and driven by some simple underlying factors that are not likely to slow down anytime soon. Specifically, the power or intelligence of an AI system can be measured roughly by multiplying together three things: (1) the quantity of chips used to train it, (2) the speed of those chips, (3) the effectiveness of the algorithms used to train it. The quantity of chips used to train a model is increasing by 2x-5x per year. Speed of chips is increasing by 2x every 1-2 years. And algorithmic efficiency is increasing by roughly 2x per year. These compound with each other to produce a staggering rate of progress. Things that seemed impossible for AI systems to do, often become routine and taken for granted a couple years later: for example, two years ago the idea of an AI system telling a good joke was considered absurd, whereas today's chatbots do it frequently.

I was one of the researchers who first documented this trend of smooth, rapid improvement when I worked at OpenAI back in 2018. Since then I have seen it borne out many times as the frontier of AI advances.

A key implication of all of this is that it's important to *skate to where the puck is going* – to set (or at least attempt to set) policy for where the technology will be in 2-3 years, which may be radically different from where it is right now.

**Short-Term, Medium-Term, and Long-Term Risks**

With the fast pace of progress in mind, we can think of AI risks as falling into three buckets:

- **Short-term** risks are those present in current AI systems or that imminently will be present. This includes concerns like privacy, copyright issues, bias and fairness in the model's outputs, factual accuracy, and the potential to generate misinformation or propaganda.
- **Medium-term** risks are those we will face in two to three years. In that time period, Anthropic's projections suggest that AI systems may become much better at science and engineering, to the point where they could be misused to cause large-scale destruction, particularly in the domain of biology. This rapid growth in science and engineering skills could also change the balance of power between nations.
- **Long-term** risks relate to where AI is ultimately going. At present, most AI systems are passive and merely converse with users, but as AI systems gain more and more autonomy and ability to directly manipulate the external world, we may face increasing challenges in controlling them. There is a spectrum of problems we could face related to this, at the extreme end of which is concerns about whether a sufficiently powerful AI,

without appropriate safeguards, could be a threat to humanity as a whole – referred to as *existential risk*. Left unchecked, highly autonomous, intelligent systems could also be misused or simply make catastrophic mistakes.

Note that there are some concerns, like AI's effects on employment, that don't fit neatly in one bucket and probably take on a different form in each time period.

Short-term risks are in the news every day and are certainly important. I expect we'll have many opportunities to discuss these in this hearing, and much of Anthropic's research applies immediately to those risks: our constitutional AI principles include attempts to reduce bias, increase factual accuracy, and show respect for privacy, copyright, and child safety. Our red-teaming is designed to reduce a wide range of these risks, and we have also published papers on using AI systems to correct their own biases and mistakes. There are a number of proposals already being considered by the Congress relating to these risks.

The long-term risks might sound like science fiction, but I believe they are at least potentially real. Along with the CEOs of other major AI companies and a number of prominent AI academics (including my co-witnesses Professors Russell and Bengio) I have signed a statement emphasizing that these risks are a challenge humanity should not neglect. Anthropic has developed evaluations designed to measure precursors of these risks and submitted its models to independent evaluators. And our work on interpretability is also designed to someday help with long-term risks. However, the abstract and distant nature of long-term risks makes them hard to approach from a policy perspective: our view is that it may be best to approach them indirectly by addressing more imminent risks that serve as practice for them.

The *medium-term risks* are where I would most like to draw the subcommittee's attention. Simply put, a straightforward extrapolation of the pace of progress suggests that, in 2-3 years, AI systems may facilitate extraordinary insights in broad swaths of many science and engineering disciplines. This will cause a revolution in technology and scientific discovery, but also greatly widen the set of people who can wreak havoc. In particular, I am concerned that AI systems could be misused on a grand scale in the domains of cybersecurity, nuclear technology, chemistry, and especially biology. I will provide a high-level summary of research Anthropic has conducted in the domain of biology which may help to shed light on these concerns.

**AI, Biology, and National Security**

Over the last six months, Anthropic, working in collaboration with world-class biosecurity experts, has conducted an intensive study of the potential for LLMs to contribute to the misuse of biology. I will describe our findings at a very coarse level of detail here. I am happy to give a more detailed private briefing to any Senator interested in this topic. In addition, we have recently briefed a number of officials within the US government and private research institutes, all of whom found our results disquieting. Note also that RAND Corporation CEO Jason Matheny mentioned some similar concerns in his March 8th, 2023 Senate Testimony.

Today, certain steps in the use of biology to create harm involve knowledge that cannot be found on Google or in textbooks and requires a high level of specialized expertise. The question we and our collaborators studied is whether current AI systems are capable of filling in some of the more-difficult steps in these production processes. We found that today's AI systems can fill in *some* of these steps, but incompletely and unreliably – they are showing the first, nascent signs of risk. **However, a straightforward extrapolation of today's systems to those we expect to see in 2-3 years suggests a substantial risk that AI systems will be able to fill in all the missing pieces, if appropriate guardrails and mitigations are not put in place.** This could greatly widen the range of actors with the technical capability to conduct a large-scale biological attack.

After discovering this risk, Anthropic has introduced mitigations to ensure our currently deployed AI system is not misused in this way. For example, focusing specifically on biology, we fine tuned models with constitutional AI to make them less likely to respond to potentially harmful requests for information. We also built safety systems to identify and disrupt users seeking to violate our Acceptable Use Policy.

Our takeaway from this work is that this kind of red teaming is difficult, but essential, and particularly important right now. We think more red teaming work should happen relatively urgently in areas of national security. It would be natural for third parties and government to take a lead here, especially in domains where they have specialized expertise.

Further, labs could share both risks and risk mitigations they discover. It seems likely that many valuable mitigations will also be straightforward to implement. To this end, we are piloting a responsible disclosure process with other labs, where we will work on short-term risks at the same time as looking ahead to future ones. However, we are concerned that, even if Anthropic and other responsible developers succeed in mitigating these risks, not every actor will behave responsibly. Bad actors could build their own AI from scratch, steal it from the servers of an AI company, or repurpose open-source models if powerful enough open-source models become available.

While biology is one of our greatest concerns, we suspect that similar misuse may be possible in the cyber, chemical, and nuclear domains.

**Policy Recommendations**

In our view these concerns merit an urgent policy response. The ideal policy response would address not just the specific risks we've identified above, but would at the same time provide a framework for addressing as many other risks as possible – without, of course, hampering innovation more than is necessary. We recommend three broad classes of policies:

- First, the U.S. must **secure the AI supply chain**, in order to maintain its lead while keeping these technologies out of the hands of bad actors. This supply chain runs all the way from semiconductor manufacturing equipment to AI models stored on the

servers of companies like ours. A number of governments have taken steps in this regard. Specifically, the critical supply chain includes:
    - Semiconductor manufacturing equipment, such as lithography machines.
    - Chips used for training AI systems, such as GPUs.
    - Trained AI systems, which are vulnerable to "export" through cybertheft or uncontrolled release.
        - Companies such as Anthropic and others developing frontier AI systems should have to comply with stringent cybersecurity standards in how they store their AI systems. We have shared with the U.S. government and other labs our views of appropriate cybersecurity best practices, and are moving to implement these practices ourselves.
- Second, we recommend a **"testing and auditing regime" for new and more powerful models**. Similar to cars or airplanes, we should consider the AI models of the near future to be powerful machines which possess great utility, but that can be lethal if designed badly or misused. New AI models should have to pass a rigorous battery of safety tests both during development and before being released to the public or to customers.
    - National security risks such as misuse of biology, cybersystems, or radiological materials should have top priority in testing due to the mix of imminence and severity of threat.
    - However, the tests could also cover other concerns such as bias, potential to create misinformation, privacy, child safety, and respect for copyright.
    - Similarly, the tests could measure the capacity for autonomous systems to escape control, beginning to get a handle on the risks of future systems. There are already nonprofit organizations, such as the Alignment Research Center, attempting to develop such tests.
    - It is important that testing and auditing happen at regular checkpoints during the process of training powerful models to identify potentially dangerous capabilities or other risks so that they can be mitigated before training progresses too far.
    - The recent voluntary commitments announced by the White House commit some companies (including Anthropic) to do this type of testing, but legislation could go further by mandating these tests for all models and requiring that they pass according to certain standards before deployment.
    - It is worth stating clearly that given the current difficulty of controlling AI systems even where safety is prioritized, there is a real possibility that these rigorous standards would lead to a substantial slowdown in AI development, and that this may be a necessary outcome. Ideally, however, the standards would catalyze innovation in safety rather than slowing progress, as companies race to become the first company technologically capable of safely deploying tomorrow's AI systems.
- Third, we should recognize that the science of testing and auditing for AI systems is in its infancy, and much less developed than it is for airplanes and automobiles. In particular, it is not currently easy to entirely understand what bad behaviors an AI system is capable of, without broadly deploying it to users. Thus, it is important to **fund both**

**measurement and research on measurement**, to ensure a testing and auditing regime is actually effective.

- Our suggestion for the agency to oversee this process is NIST, whose mandate focuses explicitly on measurement and evaluation. However many other agencies could also contribute expertise and structure to this work.
- Anthropic has been a vocal supporter of the proposed National AI Research Resource (NAIRR). The NAIRR could, among other purposes, be used to fund research on measurement, evaluation, and testing, and could do so in the public interest rather than tied to a corporation.

The three directions above are synergistic: responsible supply chain policies help give America enough breathing room to impose rigorous standards on our own companies, without ceding our national lead. Funding measurement in turn makes these rigorous standards meaningful.

In conclusion, it is essential that we mitigate the grave national security risks presented by near-future AI systems, while also maintaining our lead in this critical technology and reaping the benefits of its advancement.