# Written Statement of Andrew Ng
# Before the U.S. Senate AI Insight Forum

## AI is a General Purpose Technology with numerous beneficial uses and vastly overhyped fears of catastrophe

**AI is a General Purpose Technology with numerous use cases**

AI refers to a large class of software that helps to make decisions, often by using mathematical equations to process data. When thinking about risks of AI, it is important to recognize that it is a General Purpose Technology, similar to electricity and the internet. If someone were to ask what electricity is good for, it is almost hard to answer because there are so many use cases. Today's AI is useful for answering questions, tutoring students, acting as a brainstorming partner, copyediting, customer support, generating art, self-driving cars, deciding whether to underwrite a loan, detecting disease, and many more applications.

The generality of AI means that -- like electricity -- some can use it for nefarious purposes. But General Purpose Technologies such as electricity, the internet, and the internal combustion engine have been responsible for some of the greatest increases in people's wellbeing. As in those cases, government has an important role to play to enable AI to lift up citizens while minimizing harmful applications.

However, it is important to differentiate regulating applications (which we need) vs. regulating the technology (which is ill-advised).

**We need good regulations for AI applications**

AI technology is used in applications in healthcare, underwriting, self-driving, social media, and other sectors. With some applications, there are risks of significant harm. We want:

- Medical devices to be safe
- Underwriting software to be fair, and not discriminate based on protected characteristics
- Self-driving cars to be safe
- Social media to be governed in a way that respects freedom of speech but also does not subject us to foreign actors' disinformation campaigns

When we think about specific AI applications, we can figure out what outcomes we do want (such as improved healthcare) and do not want (such as medical products that make false claims) and regulate accordingly.

A fundamental distinction in decisions about how to regulate AI is between *applications* vs. *technology*.

Nikola Tesla's invention of the AC (alternating current) electric motor was a *technology*. When this technology is incorporated into either a blender or an electric car, the blender or car is an application. Electric motors are useful for so many things it is hard to effectively regulate them separately from thinking about concrete use cases. But when we look at blenders and electric cars, we can systematically identify benefits and risks and work to enable the benefits while limiting risks.

Whereas motors help us with physical work, AI helps us with intellectual work.

In the case of AI, engineers and scientists will typically write software and have it learn from a lot of data. This AI system may live in a company's datacenter and be subject to testing and experimentation, but it is not yet made available to any end-user. This is *AI technology* -- essentially a piece of math that can be used for many

different applications. When engineers then use this technology to build a piece of software for a particular purpose, such as medical diagnosis, it then becomes an application.

**Regulating large AI models makes as much sense as regulating high horsepower motors**

Some regulatory proposals, including a recent White House executive order, use an AI model's size (or, more precisely, the amount of computation used to develop the model) to determine risk. This is a flawed approach.

It is true that a motor capable of delivering 100 horsepower (HP), suitable for use in an electric car, can do more damage than one capable of 2 HP (suitable for a powerful blender). However, HP is a flawed metric for the risk of a motor. Indeed, either a 100 HP a 2 HP motor can be used:

- To create a dangerous medical device
- To power machinery in a lab creating a bioweapon
- To power an armed drone or steer a nuclear weapon

It makes sense to regulate medical devices, biological products, and weapons. Not motors.

Similarly, a small AI model or a large AI model can be used:

- To give bad medical advice
- To generate disinformation
- To control weapons

Some organizations have made naive statements asserting that large AI models are inherently more dangerous. This makes as much sense as saying "Large AC motors are inherently more dangerous." While it is true that one can build more dangerous things with large AC motors than small ones, it is ultimately the applications, rather than the size of the AC motor, that we should be more worried about.

**A tiered approach to risk: Applications with large reach have more potential for harm**

We take a tiered approach to regulating many activities. For example, OSHA places more stringent requirements on large than small employers. This balances the twin goals of protecting workers while not overly burdening small businesses that don't have as many resources to meet complex compliance requirements, including ones that might grow into large employers some day.

Similarly, we should take a tiered approach to regulating AI applications according to their degree of risk. The EU AI Act tries to take a tiered approach to risk, but doing this effectively requires that we clearly identify what is actually risky. To determine risk, the nature of the application, such as whether AI contributes to a medical device, which carries a life-or-death risk, or a chat system, which carries a risk of mis- or disinformation, is one factor.

An additional significant factor is reach. For example, a social media company that reaches 100 million users creates significantly more risk (of disinformation and misinformation) than a small message board reaching only 100 users. Similarly, a Generative AI company chatting with 100 million users has much greater risk than a research project testing on 100 users. We should demand greater transparency from organizations with a large reach. Fortunately, institutions that have a lot of users are also much more likely to have the resources to handle more complex compliance requirements, and so a regulatory approach that governs applications increasingly as their reach expands can bring significant protection to citizens without hampering innovation from a startup that has only 100 users today but aspires to compete with the big tech companies in the future.

**There is significant hype about AI harms**

Our media environment amplifies sensationalist statements about hypothetical AI harms, while drawing little attention to thoughtful, balanced statements about actual AI risks. For example, a statement making a highly questionable analogy between AI and nuclear weapons received ample media coverage, despite not presenting any credible reasons why AI (which helps us make better decisions) should be considered similar to nuclear weapons (which blow up cities).

While some people who promote such fears are sincere about their worries, there are numerous incentives to create hype around AI fear:

- Large companies, including ones that would rather not have to compete with open-source software, are interested in hyping up fear to create regulations -- for example requiring licensing or mandating burdensome reporting requirements for AI software -- that would "pull up the ladder" to make it difficult for new entrants, thus locking in their market.
- Non-profits have a financial incentive to create phantoms so that they can raise funds to fight the phantoms they conjured.
- Many individuals have gained significant visibility through loudly hyping up AI harms, which translates to speaker fees and other benefits.

Previous waves of general purpose technologies came with worries that turned out to be unjustified. The Pessimists Archive website documents many such examples. Of course, there have been some technologies such as the internet or social media whose benefits also came with significant downsides. But we should make sure not to stifle innovation -- including specifically open-source software, which has been a key mechanism for technological innovation -- through overregulation.

**The hype itself about AI risks is causing real harm**

I have spoken with many of the leading voices who are worried about AI catastrophic risks, including extinction risk. In the case of extinction risk, I just don't get it -- I don't see any plausible path for AI to lead to human extinction. Such sensationalist statements are already leading to harm:

- There are students who are discouraged from entering AI because they don't want to contribute to human extinction. It is a loss of opportunity for the student and for society.
- Congressional and White House leadership have limited bandwidth for dealing with AI issues and sensationalist worries about catastrophic risks may distract them from paying attention to actually risky AI products and doing the work to safeguard us from them.

Most arguments that AI could cause human extinction boil down to "it could happen." Trying to prove that it can't is akin to proving a negative. I am no more able to prove that AI won't kill us all than prove that radio waves emitted from our planet won't cause space aliens to find us and wipe us out.

As Nobel Laureate Daniel Kahneman writes in *Thinking, Fast and Slow*:

> If you are asked about the probability of a tsunami hitting California within the next thirty years, the images that come to your mind are likely to be images of tsunamis…. You will be prone to overestimate the probability of a disaster.

We know from behavioral economics that people tend to overweight very low-probability events. If a catastrophe has only a 0.0001% chance, someone asked to estimate it is likely to give a higher number. It is not surprising that many people -- even some AI scientists -- overestimate the risk of AI catastrophes. We need a better approach to assess actual AI risks other than simply appealing to authority and taking the average of people's opinions.

**The need to study and understand actual AI risks**

There have been several hypothesized AI catastrophic harms, including:

- AI "taking over", leading to human extinction
- Cybersecurity risks created through using AI to identify security vulnerabilities
- Bad actors, including enemy states, using AI to create a bioweapon

For many of these, it is possible to "red-team" out concrete scenarios and carry out back-of-the-envelope estimates (also called Fermi estimates) that will result in a more sensible estimate of the risk. The key to constructing such an estimate is to:

1. List the key sequence of steps needed for a particular catastrophe to occur
2. Estimate the probability of each step
3. Multiply the probabilities to form an estimate of the overall risk

Example: Hypothetical estimate of risk of AI "taking over", leading to human extinction

As a hypothetical example of a calculation to estimate the risk of AI "taking over" and causing human extinction over the next 100 years, a plausible scenario might be:

- **One of the world's top AI systems goes rogue, is "misaligned" and either deliberately or accidentally picks a goal that involves wiping out humanity.** Say, AI is asked to reduce carbon emissions, and decides wiping out humans is the best way to do that. Today's advanced AI systems are already smart enough to know that the default is that we want it to obey laws and not create harm. Indeed, as Arvind Narayanan and Sayash Kapoor point out, today's AI is already very good at avoiding accidental harms (as opposed to malicious, expert adversary directed harms), My estimate: less than 0.1% chance (or 1/1000).
- **Other leading AI systems do not identify the rogue actor and raise the alarm/act to stop the rogue one.** Considering the growing capabilities in companies, academia and across society of identifying AI risks, this seems unlikely to me. My estimate: less than 1% chance (or 1/100).
- **This rogue AI system gains the ability (perhaps access to nuclear weapons, or skill at manipulating people into using such weapons) to wipe out humanity.** Human extinction -- wiping out every human -- within 100 years likely requires making Earth inhospitable to human life. Considering the safeguards around tools that can do that, such as nuclear weapons, this seems very hard even for a future AI. My estimate: less than 1% chance (or 1/100).

If we multiply these numbers together, we end up with 0.1% * 1% * 1% = 1/10,000,000 chance.

Perhaps it is scary that there is any non-zero chance of human extinction. But if there's a risk that over any 100 year period leads to a 1/10,000,000 chance of extinction, this means statistically we would expect this particular cause to lead to human extinction in around 1,000,000,000 (one billion) years.

This calculation is intended to be a hypothetical one to illustrate the process of constructing a Fermi estimate rather than give a definitive answer, and I have made debatable assumptions (such as that if an AI wipes out humanity that it will be one of the top systems). A more rigorous study of the sequence of steps and estimate of the risk of each of the steps would lead to a more accurate estimate. Regardless, I believe the risk of this scenario is vanishingly small. A similar exercise, carried out through detailed study, for other risk scenarios would help us to better quantify and address any catastrophic risks.

Cybersecurity risks created through using AI to identify security vulnerabilities

There is also fear that AI could create cybersecurity risks by being very good at exposing security vulnerabilities. But if vulnerabilities get exposed more easily, the advantage lies with the defender. This is why

some of the most secure software is open-source software -- which anyone can look at to identify vulnerabilities -- because such vulnerabilities get patched quickly. I expect open-sourcing AI will also be a good way to ensure it is safe.

One realistic risk is if the attacker and defender have asymmetric capabilities. For example, a nation state might marshal advanced AI to attack a regional bank which does not have comparable capabilities for identifying vulnerabilities. In this case, the attacker might identify vulnerabilities that the defender cannot. Fortunately, by investing in AI-enabled defenses, we can make software more secure and help all parties in our nation better identify and patch their own vulnerabilities. The government has an important role in ensuring timely dissemination of security best practices, even as such best practices evolve.

AI enabled bioweapons

If you ask ChatGPT how to win an Olympic medal, it will give helpful suggestions such as find a good coach, pick the right sport, and train hard. Nonetheless, winning an Olympic medal remains really hard! Similarly, AI could offer suggestions to someone wishing to create a bioweapon, but the lab work, experimentation, and multiple steps needed to create and disseminate a bioweapon remain very difficult. I have spoken with subject matter experts (biologists) on this topic, the majority of whom seem to think that even with AI, creating a bioweapon is hard. But a more detailed understanding of the plausible scenarios -- specifically, listing the sequence of steps that would have to happen to lead to this catastrophe -- would arm us to better understand this risk as well as put in place reasonable regulations to choke off paths to this outcome.

Further, we might have an early warning signal for when bioweapons are becoming easier to build: The early warning will be when pharmaceutical companies are able to inexpensively and quickly create numerous cures. Because before AI makes it possible for a low-resource terrorist organization (or malevolent individual) to create highly capable organisms for harm, it is likely that it will first make it possible for highly resourced pharmaceutical companies to create many new biologics and drugs that benefit us.

**Funding NIST, Department of Energy, and other agencies to understand and quantify risks**

There is important work to be done to understand concrete AI risk scenarios and estimate their probabilities. This is true both for catastrophic harms and for less catastrophic ones, such as misinformation, amplifying hate, bias, and concentration of power.

Several government agencies, including NIST and the Department of Energy, have significant technical expertise in AI and large-scale computing similar to that used to build advanced AI. (For some risks, collaborating with other agencies, such as HHS and DHS in the case of bio risks, would be appropriate.) Teams in such agencies, and in academia, are well equipped to brainstorm a comprehensive list of risks, lay out sequences of steps that would lead to hypothetical harms, and quantify their odds.

However, salaries of AI engineers are high, and we must ensure that such organizations are well funded to attract the AI talent needed to study and quantify AI risks. Consequently, a key step to mitigating risks is to increase funding to these organizations and to mandate that they develop realistic AI risk scenarios and quantify the chance of each, and filter out which worries are actually realistic.

**Mandate AI Transparency to enable regulators, media, academia and civic society to spot problems**

As Senator Blumenthal recently said, “Congress failed to meet the moment on social media. Now we have the obligation to do it on AI before the threats and the risks become real.” With social media, we found out many of

the problems through whistleblowers (such as Christopher Wylie in the case of Cambridge Analytica), investigative journalism, academic studies, and luck.

With the rise of AI, there are likely problems lurking in myriad applications -- for example, are AI systems that simulate a boyfriend/girlfriend emotionally manipulating people to fall in love with them, then using that "relationship" to profit? -- that we might never learn about without a dose of luck. We need to do better.

In addition to providing resources to identify risks, we also need better transparency into the largest platforms -- the ones that reach many users, and thus have the greatest potential for significant harm -- to understand their impact. By giving government agencies such as the NSF as well as academic organizations and journalists the ability to obtain relevant information from large platforms, we open a much-needed window into what these products are doing and increase the odds of spotting problems quickly. Further, as I explain elsewhere, such information can be released without compromising individual user privacy.

Sometimes a company itself might not be aware of the harm it is creating. For example, for years, YouTube's recommendation engine was tuned to maximize user engagement. But even YouTube employees were not broadly aware of how this led their algorithm to amplify hateful content.

The proposed PATA legislation is a good step, but it should be expanded to cover not just social media companies, but any company that delivers content to a large audience. As we look forward to the next generation of AI companies, mandating transparency will help the government understand risks and the companies themselves understand their own impact.

**Bringing more intelligence to the world**

Even as we worry about hypothetical AI catastrophic risks, we must also consider how AI is a critical part of our response to known risks that aren't directly related to AI, such as climate change and pandemics. For example, large AI models such as Microsoft's ClimaX are helping us to understand weather and climate, and AI is used in numerous ways to ameliorate the harms of climate change. AI also holds promise for improving healthcare, including dramatically speeding up our response to pandemics. So even as we study how AI technology might *increase* certain catastrophic risks, we must also model how it *decreases* other risks to make sure our approaches to risk mitigation do more good than harm.

Intelligence is the power to apply knowledge and skills to make good decisions. We invest years of our lives and trillions of dollars on education -- all to develop our ability to make better decisions. It costs a lot to feed, educate, and train a wise human being: Human intelligence is expensive! That's why it is so costly to hire intelligence such as a specialist doctor to consult on a medical condition or a skilled tutor to coach your child.

Unlike human intelligence, artificial intelligence can be made cheap. AI has the potential to give every individual the ability to hire intelligence at a low cost, so you no longer need to worry about that huge bill for visiting a doctor or getting an education, and you can hire an army of smart, well-informed staff to think things through for you. And it has the potential to give society more intelligent guidance on how to approach some of our biggest problems, such as climate change and pandemics.

Yes, AI can be used for nefarious purposes. That's why we need to understand and quantify the risk of different AI applications, and put in place transparency and risk mitigation requirements for truly risky ones. But regulations that impose burdensome requirements on open-source software or on AI technology development would hamper innovation, have anti-competitive effects that contribute to an entrenchment of today's big tech companies, and slow down our ability to bring AI's benefits to everyone.