# Written Statement of Dr. Max Tegmark to the AI Insight Forum - Future of Life Institute

The Future of Life Institute President addresses the AI Insight Forum on AI innovation and provides five US policy recommendations.

By Max Tegmark

11 min. read · View original

**AI Insight Forum: Innovation**
October 24, 2023

**Written Statement of Dr. Max Tegmark**
Co-Founder and President of the Future of Life Institute
Professor of Physics at Massachusetts Institute of Technology

I first want to thank Majority Leader Schumer, the AI Caucus, and the rest of the Senators and staff who organized today's event. I am grateful for the opportunity to speak with you all, and for your diligence in understanding and addressing this critical issue.

My name is Max Tegmark, and I am a Professor of Physics at MIT's Institute for Artificial Intelligence and Fundamental Interactions and the Center for Brains, Minds and Machines. I am also the President and Co-Founder of the Future of Life Institute (FLI), an independent non-profit dedicated to realizing the benefits of emerging technologies and minimizing their potential for catastrophic harm.

Since 2014, FLI has worked closely with experts in government, industry, civil society, and academia to steer transformative technologies toward improving life through policy research, advocacy, grant-making, and educational outreach. In 2017, FLI coordinated development of the Asilomar AI Principles, one of the earliest and most influential frameworks for the governance

of AI. FLI serves as the United Nations Secretary General's designated civil society organization for recommendations on the governance of AI, and has been a leading voice in identifying principles for responsible development and use of AI for nearly a decade.

More recently, FLI made headlines by publishing an open letter calling for a six-month pause on the training of advanced AI systems more powerful than GPT-4, the state-of-the-art at the time of its publication. It was signed by more than 30,000 experts, researchers, industry figures, and other leaders, and sounded the alarm on ongoing, unchecked, and out-of-control AI development. As the Letter explained, the purpose of this pause was to allow our social and political institutions, our understanding of the capabilities and risks, and our tools for ensuring the systems are safe, to catch up as Big Tech companies continued to race ahead with the creation of increasingly powerful, and increasingly risky, systems. In other words, "powerful AI systems should be developed only once we are confident

that their effects will be positive and their risks will be manageable."

## Innovation does not require uncontrollable AI

The call for a pause was widely reported, but many headlines missed a crucial nuance, a clarification in the subsequent paragraphs key to realizing the incredible promise of this transformative technology. The letter went on to read:

> This does *not* mean a pause on AI development in general, merely a stepping back from the dangerous race to ever-larger unpredictable black-box models with emergent capabilities.

> AI research and development should be refocused on making today's powerful, state-of-the-art systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal.

It is not my position, nor is it the position of FLI, that AI is inherently bad. AI promises remarkable benefits – advances in healthcare, new avenues for scientific discovery, increased

productivity, among many more. What I am hoping to convey, however, is that we have no reason to believe vastly more complex, powerful, opaque, and uncontrollable systems are necessary to achieve these benefits. That innovation in AI, and reaping its untold benefits, does not have to mean the creation of dangerous and unpredictable systems that cannot be understood or proven safe, with the potential to cause immeasurable harm and even wipe out humanity.

AI can broadly be grouped into three categories:

- **“Narrow” AI systems** – AI systems that are designed and optimized to accomplish a specific task or to be used in a specific domain.
- **Controllable general-purpose AI systems** – AI systems that can be applied to a wide range of tasks, including some for which they were not specifically designed, with general proficiency up to or similar to the brightest human minds, and potentially exceeding the brightest human minds in some domains.
- **Uncontrollable AI systems** – Often referred to as “superintelligence,” these are AI systems that far exceed human capacity across virtually all

> cognitive tasks, and therefore by definition cannot be understood or effectively controlled by humans.

The first two categories have already yielded incredible advances in biochemistry, medicine, transportation, logistics, meteorology, and many other fields. There is nothing to suggest that these benefits have been exhausted. In fact, experts argue that with continued optimization, fine-tuning, research, and creative application, **the current generation of AI systems can effectively accomplish nearly *all* of the benefits from AI we have thus far conceived, with several decades of accelerating growth**. We do not need more powerful systems to reap these benefits.

Yet it is the stated goal of the leading AI companies to develop the third, most dangerous category of AI systems. A May 2023 blog post from OpenAI rightly points out that "it's worth considering why we are building this technology at all." In addition to some of the benefits mentioned above, the blog post justifies continued efforts to develop superintelligence by espousing that "it would be [...] difficult to stop the creation

of superintelligence" because "it's inherently part of the technological path we are on."

The executives of these companies have acknowledged that the risk of this could be catastrophic, with the legitimate potential to cause mass casualties and even human extinction. In a January 2023 interview, Sam Altman, CEO of OpenAI, said that "the bad case [...] is, like, lights out for all of us." In May 2023, Altman, along with Demis Hassabis, CEO of Google Deepmind, Dario Amodei, CEO of Anthropic, and more than 350 other executives, researchers, and engineers working on AI endorsed a statement asserting that "itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

It is important to understand that creation of these systems is not inevitable, particularly before we can establish the societal, governmental, and technical mechanisms to prepare for and protect against their risks. The race toward creating these uncontrollable AI

systems is the result of a tech sector market dynamic where prospective investment and perverse profit incentives drive reckless, runaway scaling to create the most powerful possible systems, at the expense of safety considerations. This is what “innovation” means to them.

But creating the most powerful system does not always mean creating the system that best serves the well-being of the American people. Even if we “win” the global race to develop these uncontrollable AI systems, we risk losing our social stability, security, and possibly even our species in the process. Far from ensuring geopolitical dominance, the destabilizing effect of haphazard proliferation of increasingly powerful AI systems is likely to put the United States at a substantial geopolitical disadvantage, sewing domestic discord, threatening national security, and harming quality of life. Our aspirations should instead be focused on innovation that improves our nation and our lives by ensuring that the systems we deploy are controllable, predictable, reliable, and

safe – systems that do what we want them to, and do it well.

For a cautionary example, we can look to the emergence of recommender algorithms in social media. Over the past decade, tremendous strides were made in developing more effective algorithms for recommending content based on the behavior of users. Social media in general, and these algorithms in particular, promised to facilitate interpersonal connection, social discourse, and exposure to high-quality content.

Because these systems were so powerful and yet so poorly understood, however, society was not adequately equipped to protect against their potential harms. The prioritization of engagement in recommender systems led to an unforeseen preference for content evocative of negative emotion, extreme polarization, and the promotion of sensationalized and even fabricated "news," fracturing public discourse and significantly harming mental and social health in the process. The technology was also weaponized against the

American people by our adversaries, exacerbating these harms.

For uncontrollable AI systems, these types of misaligned preferences and unexpected ramifications are likely to be even more dangerous, unless adequate oversight and regulation are imposed. Much of my ongoing research at MIT seeks to advance our understanding of mechanistic interpretability, a field of study dedicated to understanding how and why these opaque systems behave the way they do. My talented students and colleagues have made incredible strides in this endeavor, but there is still much work to be done before we can reliably understand and predict the behavior of today's most advanced AI systems, let alone potential systems that can operate far beyond human cognitive performance.

## AI innovation depends on regulation and oversight

Though AI may be technically complex, Congress has extensive experience putting in place the necessary governance to mitigate risks from new technologies without foreclosing their

benefits. In establishing the Federal Aviation Administration, you have facilitated convenient air travel, while ensuring that airplanes are safe and reliable. In establishing the Food and Drug Administration, you have cultivated the world's leading pharmaceutical industry, treating ailments previously thought untreatable, while ensuring that the medicine we take is safe and will not cause undue harm.

The same can and should be done for AI. In order to harness the benefits of AI and minimize its risks, it is essential that we invest in further improving our understanding of how these systems work, and that we put in place the oversight and regulation necessary to ensure that if these systems are created and deployed, that they will be safe, ethical, reliable, and beneficial.

Regulation is often framed as an obstacle to innovation. But history has shown that failure to adequately regulate industries that pose catastrophic risk can be a far greater obstacle to technological progress. In 1979, the Three Mile Island nuclear reactor suffered a partial

meltdown resulting from a mechanical failure, compounded by inadequate training and safety procedures among plant operators and management.

Had the nuclear energy industry been subject to sufficient oversight for quality assurance of materials, robust auditing for safe operating conditions, and required training standards for emergency response procedures, the crisis could likely have been avoided. In fact, subsequent investigations showed that engineers from Babcock & Wilcox, the developers of the defective mechanism, had identified the design issue that caused the meltdown prior to the event, but failed to notify customers.

The result of this disaster was a near-complete shuttering of the American nuclear energy industry. The catastrophe fueled ardent anti-nuclear sentiment among the general public, and encouraged reactionary measures that made development of new nuclear power plants costly and infeasible. Following the incident at Three Mile Island, no new nuclear power plants were authorized for construction in the United States for over

30 years, foreclosing an abundant source of clean energy, squandering a promising opportunity for American energy independence, and significantly hampering innovation in the nuclear sector.

We cannot afford to risk a similar outcome with AI. The promise is too great. By immediately implementing proactive, meaningful regulation of the AI industry, we can reduce the probability of a Three Mile Island-like catastrophe, and safeguard the future of American AI innovation.

## Recommendations

To foster sustained innovation that improves our lives and strengthens our economy, the federal government should take urgent steps by enacting the following measures:

1. Protect against catastrophes that could derail innovation, and ensure that powerful systems are developed and deployed only if they will safely benefit the general public. To do so, we must require that highly-capable general purpose AI systems, and narrow AI systems intended for use in high-risk applications such as critical infrastructure, receive independent

audits and licensure before deployment. Importantly, the burden of proving suitability for deployment should fall on the developer of the system, and if such proof cannot be provided, the system should not be deployed. This means approval and licensure for development of uncontrollable AI should not be granted at all, at least until we can be absolutely certain that we have established sufficient protocols for training and deployment to keep these systems in check.

Auditing should include pre-training evaluation of safety and security protocols, and rigorous pre-deployment assessment of risk, reliability, and ethical considerations to ensure that the system does not present an undue risk to the well-being of individuals or society, and that the expected benefits of deployment outweigh the risks and harmful side effects. These assessments should include evaluation of potential risk from publishing the system's model weights – an irreversible act that makes controlling the system and derivative systems virtually impossible – and provide requisite limitations on publication of and access to model weights as a condition of licensure. The process should also include continued monitoring and reporting of potential safety, security, and ethical concerns throughout the lifetime of the AI system. This will help identify and correct emerging

and unforeseen risks, similar to the pharmacovigilance requirements imposed by the FDA.

2. Develop and mandate rigorous cybersecurity standards that must be met by developers of advanced AI to avoid the potential compromise of American intellectual property, and prevent the use of our most powerful systems against us. To enforce these standards, the federal government should also require registration when acquiring or leasing access to large amounts of computational hardware, as well as when conducting large training runs. This would facilitate monitoring of proliferation of these systems, and enhance preparedness to respond in the event of an incident.
3. Establish a centralized federal authority responsible for monitoring, evaluating, and regulating general-purpose AI systems, and advising other agencies on activities related to AI within their respective jurisdictions. In many cases, existing regulatory frameworks may be sufficient, or require only minor adjustments, to be applicable to narrow AI systems within specific sectors (e.g. financial sector, healthcare, education, employment, etc.). Advanced general-purpose AI systems, on the other hand, cut across several jurisdictional domains, present unique risks and novel capabilities, and are not adequately addressed by existing, domain-specific regulations or authorities. The centralized body

would increase the efficiency of regulating these systems, and help to coordinate responses in the event of an emergency caused by an AI system.

4. Subject developers of advanced general-purpose AI systems (i.e. those with broad, unpredictable, and emergent capabilities) to liability for harms caused by their systems. This includes clarifying that Section 230 of the Communications Decency Act does not apply to content generated by AI systems, even if a third-party provided the prompt to generate that content. This would incentivize caution and responsibility in the design of advanced AI systems, aligning profit motives with the safety and security of the general public to further protect against catastrophes that could derail AI innovation.
5. Increase federal funding for research and development into technical AI safety, reliable assessments and benchmarks for evaluating and quantifying risks from advanced AI systems, and countermeasures for identifying and mitigating harms that emerge from misuse, malicious use, or unforeseen behavior of advanced AI systems. This will allow our tools for assessing and enhancing the safety of systems to keep pace with advancements in the capabilities of those systems, and will present new opportunities for innovating systems better aligned with the public interest.

## Innovation is what is best, not what is biggest

I have no doubt there is consensus among those participating in this Forum, whether from government, industry, civil society, or academia, that the best path forward for AI must foster innovation, that American ingenuity should not be stifled, and that the United States should continue to act as a leader in technological progress on the global stage. That's the easy part.

The hard part is defining what exactly "innovation" means, and what type of leader we seek to be. To me, "innovation" means manifesting new ideas that make life better. When we talk about American Innovation, we are talking not just about the creation of new technology, but about how that technology helps to further democratic values and strengthen our social fabric. How it allows us to spend more time doing what we love with those we love, and keeps us safe and secure, both physically and financially.

Again, the nuance here is crucial. "Innovation" is not just the manifestation of new ideas, but also

ensuring that the realization of those ideas drives us toward a positive future. This means that a future where America is a global leader in AI innovation does not necessarily mean that we have created a more powerful system — that is, a system with more raw power, that can do more things. What it means is that we have created the systems that lead to the best possible America. Systems that are provably safe and controllable, where the benefits outweigh the risks. This future is simply not possible without robust regulation of the AI industry.