# FLI on "A Statement on AI Risk" and Next Steps

The view that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" is now mainstream, with that statement being endorsed by a who's who of AI experts and thought leaders from industry, academia, and beyond.

Although FLI did not develop this statement, we strongly support it, and believe the progress in regulating nuclear technology and synthetic biology is instructive for mitigating AI risk. **FLI therefore recommends immediate action to implement the following recommendations**.

## Recommendations:

- Akin to the Nuclear Non-Proliferation Treaty (NPT) and the Biological Weapons Convention (BWC), develop and institute international agreements to limit particularly high-risk AI proliferation and mitigate the risks of advanced AI, including track 1 diplomatic engagements between nations leading AI development, and significant contributions from non-proliferating nations that unduly bear risks of technology being developed elsewhere.
- Develop intergovernmental organizations, akin to the International Atomic Energy Agency (IAEA), to promote peaceful uses of AI while mitigating risk and ensuring guardrails are enforced.
- At the national level, establish rigorous auditing and licensing regimes, applicable to the most powerful AI systems, that place the burden of proving suitability for deployment on the developers of the system. Specifically:
    - Require pre-training auditing and documentation of a developer's sociotechnical safety and security protocols prior to conducting large training

runs, akin to the biocontainment precautions established for research and development that could pose a risk to biosafety.

    - Similar to the Food and Drug Administration's (FDA) approval process for the introduction of new pharmaceuticals to the market, require the developer of an AI system above a specified capability threshold to obtain prior approval for the deployment of that system by providing evidence sufficient to demonstrate that the system does not present an undue risk to the wellbeing of individuals, communities, or society, and that the expected benefits of deployment outweigh risks and harmful side effects.
    - After approval and deployment, require continued monitoring of potential safety, security, and ethical risks to identify and correct emerging and unforeseen risks throughout the lifetime of the AI system, similar to pharmacovigilance requirements imposed by the FDA.

- Prohibit the open-source publication of the most powerful AI systems unless particularly rigorous safety and ethics requirements are met, akin to constraints on the publication of "dual-use research of concern" in biological sciences and nuclear domains.
- Pause the development of extremely powerful AI systems that significantly exceed the current state-of-the-art for large, general-purpose AI systems.

The success of these actions is neither impossible nor unprecedented: the last decades have seen successful projects at the national and international levels to avert major risks presented by nuclear technology and synthetic biology, all without stifling the innovative spirit and progress of academia and industry. International cooperation has led to, among other things, adoption of the NPT and establishment of the IAEA, which have mitigated the development and proliferation of dangerous nuclear weapons and encouraged more equitable distribution of peaceful nuclear technology. Both of these achievements came during the height of the Cold War, when the United States, the USSR, and many others prudently recognized that geopolitical competition should not be prioritized over humanity's continued existence.

Only five years after the NPT went into effect, the BWC came into force, similarly establishing strong international norms against the development and use of biological weapons, encouraging peaceful innovation in bioengineering, and ensuring international cooperation in responding to dangers resulting from violation of those norms. Domestically, the United States adopted federal regulations requiring extreme caution in the conduct of research and when storing or transporting materials that pose considerable risk to biosafety. The Centers for Disease Control and Prevention (CDC) also published detailed guidance establishing biocontainment precautions commensurate to different levels of biosafety risk. These precautions are monitored and enforced at a range of levels, including through internal institutional review processes and supplementary state and local laws. Analogous regulations have been adopted by nations around the world.

Not since the dawn of the nuclear age has a new technology so profoundly elevated the risk of global catastrophe. FLI's own letter called on "all AI labs to immediately pause for at least six months the training of AI systems more powerful than GPT-4." It also stated that "If such a pause cannot be enacted quickly, governments should step in and institute a moratorium."

Now, two months later – despite discussions at the White House, Senate hearings, widespread calls for regulation, public opinion strongly in favor of a pause, and an explicit agreement by the leaders of most advanced AI efforts that AI can pose an existential risk – there has been no hint of a pause, or even a slowdown. If anything, the breakneck pace of these efforts has accelerated and competition has intensified.

The governments of the world must recognize the gravity of this moment, and treat advanced AI with the care and caution it deserves. AI, if properly controlled, can usher in a very long age of abundance and human flourishing. It would be foolhardy to jeopardize this promising future by charging recklessly ahead without allowing the time necessary to keep AI safe and beneficial.