NAIAC written statement by Yoshua Bengio for the meeting of the AI Futures working group of August 3, 2023

There is no doubt that AI can have tremendous positive economic, social and environmental impact and I find plausible Stuart Russell's calculation (in his 2019 book), leading to an estimated net present value of quadrillions of dollars for the upcoming advances and deployment of AI as we approach and surpass human-level capabilities. However, these will only be possible if we also correctly identify and mitigate the risks, from the better researched and established ones like bias and discrimination to those anticipated in coming years, starting with misuse (with direct threats in an estimated timeframe of 1 to 3 years to the democratic process and national security) and up to loss of control with existential risk (maybe in as little as 5 years, with great uncertainty).

Please consult my written testimony to the Judiciary Committee of the United States Senate for a comprehensive description of the anticipated risks, including potential catastrophic outcomes, as well as my recommendations, which I summarize here:

1. **AI Governance**: Rapid adoption of highly adaptive national regulation, as well as starting and accelerating negotiations to lead to the adoption of an international regulatory instrument, and the standardization of principles to mitigate major global risks.
2. **Research**: Massive investments in research and development of two kinds:
    a. Open research on AI safety and innovative governance mechanisms, which can be achieved through accelerated and increased funding to academic labs
    b. Classified research on countermeasures against eventual rogue AIs which may be created either intentionally (as harmful tools) or unintentionally (due to loss of control).

Regarding recommendation 1, I would like to emphasize that the immediate and primary action should be to engage in discussions with China, with the objective to converge on a set of principles and standards for managing and monitoring AI risks (and commitment to avoid using powerful AI in military attacks or destabilizing attempts to each other's government and economy). These principles should include an agreement on the process to revise the agreed-upon standards as research and development progress (with objective 2 above, as well as from industry-driven efforts) and we better understand the risk scenarios and mitigation options. Once China and the US agree on such principles, it will likely be easier to enlarge the conversation to include a broader set of countries and institutions, including G7 countries and other strategic countries - both like-minded and not -, as well as the international community, which could take place via international organizations such as the UN, the OECD and others. An important objective of the broader circle is to make sure that countries which do not have current AI capabilities (including the Global South) will establish the proposed national legislation and agree to international monitoring, because the Internet, computer viruses and biological viruses are not much bound by national borders. An appeal for Global South countries to sign such a treaty should be that they reap some of the benefits of the AI advances (e.g., in helping with the UN SDGs).

Regarding recommendation 2 and countermeasures against eventual rogue AIs, I have written an article (not yet published but submitted to The Journal of Democracy) which elaborates a specific proposal. In summary, I provide motivations and rationales for a decentralized, coordinated and multilateral network of research labs focusing on the development of AI and national security methodologies to enable a coalition of democratic countries to prepare countermeasures against eventual rogue AIs. I argue why these labs should be non-profit and independent while being mostly funded by governments. They would operate similarly to academic labs, except that they would be working on classified research shared only within the network and would have the financial resources to pool the high-price computational infrastructure and attract the talent needed for this work. They should closely coordinate with the national security arms of the countries in this coalition for two reasons: to help better define and understand the threats and vulnerabilities as well as to deploy the required methodologies - such as cybersecurity - across these countries. Bringing multiple like-minded countries (maybe Five Eyes) on this has several advantages, including being able to tap into AI talent outside the US, increasing the scientific diversity and effectiveness of the research, facilitating the deployment of proposed solutions across a group of countries and making the whole effort more robust to democratic downturns in one of the participating nations.