Written Statement of

Professor Yoshua Bengio
Full professor of Computer Sciences at University of Montreal,
Founder and Scientific Director of Mila - Quebec AI Institute
2018 Co-recipient of the AM Turing Award

Presented before the U.S. Senate Forum on AI Insight Regarding Risk, Alignment, and Guarding Against Doomsday Scenarios

December 6, 2023

Thanks to Valérie Pisano, Michael Cohen, Daniel Privitera, Sören Mindermann, Jon Menaster, Marc-Antoine Guérard, Benjamin Prud'homme, Niki Howe, Alan Chan, Aris Richardson, Noam Kolt, Ashwin Acharya and Ludovic Soucisse for their feedback.

Executive summary

Humanity is on a trajectory of accelerating advances in Artificial Intelligence (AI). In 2019, the most advanced model was GPT-2, a model that could not reliably count to ten. Only four years later, similar but larger AI systems also based on Deep Learning can write software and advise on intellectual topics. Tech companies are now engaged in a race to create Artificial General Intelligence (AGI): generalist and autonomous systems that match or surpass human abilities in most or all knowledge work. Three winners of the 2018 Turing award for deep learning (Geoffrey Hinton, Yann LeCun and myself) place the timeline for AGI in an interval ranging from a few years to a few decades. In this statement, I examine some larger-scale risks this entails, and I propose ways to mitigate risks of catastrophic outcomes.

There is a risk of losing control over AI with powerful capabilities, a risk we have yet to learn how to mitigate. If those in control of AI do not understand and manage this risk, it could jeopardize all of humanity. There are two core challenges behind AI-driven severe risks that are cause for urgent concern. *The first is **AI alignment***: No one currently knows how to create advanced AI that reliably follows the intent of its developers. Without this ability, we risk that even well-meaning actors unintentionally create AI systems with undesirable goals or vulnerabilities that can be exploited for malicious purposes. To advance undesirable goals, powerful AI systems could use strategies such as self-replication and deceptive behavior towards humans. *The second challenge is **social and political***: Even if we knew how to control AI, it could still be dangerous in the hands of those wishing to use its power for their own benefit, to obtain economic, political or military dominance. In the wrong hands, superhuman capabilities – or human-level capabilities at scale and low cost – can cause catastrophic harm. Even if we avoid the loss of control scenarios, we will need to take action to preserve democracy, whose institutions, checks and balances are all about avoiding concentration of power. And, if we develop solutions to the alignment problem, we will need to ensure that all actors adopt such precautions. We need to take steps to mitigate these risks today -- both because frontier AI labs may develop AGI soon, and because it will take time to develop the solutions and put them into place.

The primary recommendation of my statement is: ***if developers intend to build AI that is capable enough to have the potential to be catastrophically dangerous in the wrong hands or through loss of control, they must demonstrate that their system will be safe prior to full training and deployment of the AI***. Governments should keep track of such systems, with particular safety controls to detect and avoid self-replication and deceptive behavior by the AI. Governments should also require a secure one-way *off-switch* that the regulator can trigger if systems are not safe.

The second recommendation is to prepare for the emergence of dangerous AI: ***we must urgently advance AI alignment research and build aligned AI systems to help protect us***. We need to develop such systems under very strong democratic and multilateral governance, to ensure safety and avoid powerful AI systems being abused or falling into the wrong hands.

For the time being, safe AI can easily be made unsafe or rerouted for misuse. Raw intellectual capabilities can be directed toward any goal, good or bad. AI could therefore be used for dangerous purposes if it is in the wrong hands, or if the AI harbors a nefarious goal. Current safety protections seem to be easily undone, e.g. the safety protections designed to avoid misuse of Meta's Llama 2 were removed with a small effort and only a few hundred dollars.

Power concentration, misuse, disinformation and national security risks

Dangerous uses of AI are already beginning to become more prominent in our society, before we reach AGI. We are already seeing concrete harms from misuse (e.g., AI being used to generate deepfake nude images of teenagers and vast quantities of hate speech). Increases in AI capability are likely to yield correspondingly greater harm because AI is dual-use. An AI could bring catastrophic harm even if it does not surpass humans on every intellectual task. For example, an AI that could design a lethal and highly transmissible virus would be very dangerous even if it struggled with other tasks. Similarly, an AI that is strong at convincing people, perhaps after extensive practice on social media, could be used to influence political opinions and destabilize democracies, even if it lacked robotics abilities or scientific knowledge. These risks to our democracy and society arise from a common risk factor: the concentration of power. A single very capable AI could wield significant power, and a single organization could wield such power if they avoid losing control of the AI.

Intentional and unintentional loss of control scenarios

Companies have set their sights on building *autonomous* AI: systems that can plan, act, and pursue long-term goals. Advanced autonomous AI will pose control challenges. Control of software has been a problem since the development of computer worms. But AI is making progress in capabilities such as hacking, persuasion, and strategic planning. There is a limit to how confident we can be of dangers that have not been reproduced in a lab and studied empirically, but all academic work I am aware of on this topic points in the direction that catastrophic outcomes are a distinct possibility, not a remote one.

Autonomous AI systems are goal-driven and there are many reasons why they could end up with goals that lead to harm. Notably, this need not involve "hating" humans or anything similar. I will outline three reasons.

- First, the simplest scenario that may lead to a loss of control is if a human intentionally instructs a powerful autonomous AI to make self-preservation its primary goal. Some AI researchers publicly stated their desire to see humanity replaced by super-capable AIs, arguing that the supreme value is intelligence and that it would only be a natural succession. I consider that intentionally acting towards such a catastrophic goal for humanity should be criminal.
- Second, no one currently knows how to *reliably* embed AI systems with desirable goals, including safety constraints. For example, researchers found a particular sequence of characters to type before instructing ChatGPT, with the model then answering any question without constraints, for example giving instructions to build weapons.

- The third reason is that we do not know how to formally state human moral judgements. Instead, training of Frontier AI models typically penalizes models for bad behavior and rewards desired behavior. So far we cannot verify exactly which goals this process embeds or not in the system. A distinct possibility is that an advanced system adopts reward as a goal in itself, and takes control from the operators to provide its own rewards. This is known as reward hacking. Theoretical results show that it happens if an AI trained with rewards is capable enough.

To summarize, there are many reasons AI could end up with undesirable goals. Each of them may put the system in conflict with humanity and give the system a reason to preserve itself despite human attempts to intervene.

Advanced autonomous AI systems pursuing undesirable goals may become hard to control. To advance undesirable goals, they could use unacceptable strategies like gaining human trust, acquiring resources, using deception to influence decision-makers, and forming coalitions with humans and other AIs. AI systems are already widely used for programming. To avoid being shut down, they may copy themselves globally, like computer worms, and insert security vulnerabilities to control key systems like communication networks, financial systems, supply chains, and autonomous weapons. AI systems may also engineer more advanced AI, to better achieve their goals. Many of the leading AI academics have pointed out these problems, supported by mathematical and empirical results. We could mitigate these risks of AI adopting dangerous strategies by properly aligning their objectives to prioritize human safety over self-preservation and control, similar to Asimov's laws of robotics. Unfortunately, the technology to effectively program this level of AI alignment is not yet developed.

We may also lose control by gradually handing it over. As AI systems become faster and more cost-effective than humans, organizations may increasingly rely on AI systems instead of humans when making decisions, to keep up with competitors and adversaries who do the same. This could lead to widespread deployment of AI systems in critical societal roles, with less human oversight, due to the cost and effort involved in verifying AI decisions and goals.

Insufficiently reassuring arguments against the catastrophic scenarios

There is uncertainty about the timeline and plausibility of the above catastrophic scenarios, but the consequences could be drastic, which means that decision-makers should act to mitigate them. Society is often well-served by addressing technological risks reactively, but AI is not a typical technology: even leading AI developers admit that it has the potential for extreme harm. We do not sufficiently understand what we are doing with current AI systems, compared with almost every other field of engineering, and yet we are racing to build extremely powerful and thus potentially extremely dangerous machines. Several arguments have been made that we should not worry about such scenarios. I would like to be convinced of that, so I have collected as many of these arguments as possible, in a text available on my blog, along with the reasons why I do not find them reassuring. For example, a common suggestion is that a developer could simply unplug the AI if it exhibits dangerous behavior. But as outlined above, an advanced AI

could also be explicitly instructed to perform dangerous behaviors by unscrupulous developers, or might provide so many short-term benefits that its developers are unwilling to turn it off. An off-switch in the hands of a third-party regulator could help with these particular risks.

Regulating Frontier AI and requiring safety

My primary recommendation to avoid the above potentially catastrophic outcomes is to establish an agile regulatory framework with the following objectives (see also this document and this one for related ideas). First, the government should have the ability to stop development or deployment of advanced models. Second, for such very powerful systems with a potential for harm, authorizations should be obtained only if the developer demonstrates with appropriate scientific evidence that their system will be sufficiently safe: a potentially dangerous system should be considered unsafe until proven safe rather than vice-versa, like a new drug whose safety needs to be demonstrated, with the burden of proof on the pharmaceutical company.

An undesirable setting for the most capable and risky AI models would be one in which the government has the burden of testing and evaluating in order to uncover a potential safety problem. By contrast, a regulatory regime under which companies must prove the safety of their systems would greatly incentivize companies to invest in AI alignment and AI safety research. Since safety may never be perfectly guaranteed, having multiple layers of defense is important. Namely, it has been proposed that future Frontier AI systems should have a secured one-way off-switch that the regulator could trigger. Since an AI can be harmful across borders, it is of course essential that international agreements be put in place as soon as possible to harmonize such regulations across the globe, and maybe exclude non-signatories from the AI supply chain.

Using compute as a proxy for potentially dangerous AI

Uncontrollable AI must never be built, but we often do not know the properties of advanced AI systems before they are built. Therefore, regulators should use available indicators and proxies to decide which AI designs may be unsafe. One way suggested by President Biden's recent AI Executive Order is by requiring registration only of systems trained with extensive compute. Currently, the most capable AI systems gain those capabilities after extensive training, so regulators might be reasonably confident that low-compute systems need not be controlled as tightly. An appropriate "compute threshold [Appendix A.4]" would of course have to be adapted as AI technology becomes more efficient and AI safety better understood.

Monitoring Frontier AI systems

If regulators place controls on the development of some AIs, especially high-compute forms, they must be able to enforce them. Currently, regulators have little to no visibility into the most important inputs for advanced AI. This includes chips and data centers, but it also includes pre-trained Frontier models. Public access to a very advanced pre-trained Frontier model would likely make the (accidental) development of uncontrollable AI much easier and cheaper. Governments know where uranium is. They must know where very advanced Frontier models

are, the hardware needed to train them, and whether they are being used in ways that risk creating uncontrollable AI or national security risks. If very advanced Frontier models are made open source, no one will be able to report to the government where all the Frontier models are stored or what is being done with them, because no one will know; they will be on countless machines, making controls on compute ineffective, steadily decreasing the cost of making uncontrollable AI, and drastically reducing the possibility for democratic oversight. Advanced AI must instead be registered and monitored. As mentioned above, the recent AI Executive Order is an important first step in ensuring registration through reporting requirements but it does not impose controls. To balance the pros and cons of open source, a regulator and not the CEO of a company should decide whether a powerful Frontier model could be open-sourced.

Research in AI safety and AI alignment

To minimize the probability of loss of control, or mitigate harm if it occurs, it is imperative and urgent to figure out how to design safe AIs. This requires solving AI alignment: developing technical methods to ensure AIs behave according to our intentions and instructions, or at least do not cause major harms. This would have tremendous commercial value but it is also essential for another reason: regulation is unlikely to shield us perfectly from the emergence of powerful and dangerous AIs, either because they are in the hands of bad actors or because we have lost control and they have their own dominant self-preservation goal. To protect democracy and humanity from dangerous super-capable AIs, we may need aligned, defensive AIs to help protect us. However, absent further progress on the control problem, we must not precipitate an AI race and create the very danger we seek to avoid.

Avoiding a single point of failure and requiring strong democratic governance

Open sourcing the most capable AIs, thus making them freely available for download on the internet, could lead to catastrophic outcomes (e.g. through misuse from terrorists) and make a loss of control more likely. However, the potential for excessive power concentration would be increased if there is a single, closed AI system whose capabilities are well above others. A takeover by an authoritarian government, a mistake leading to a leak into the hands of bad actors, or a loss of control to an AGI could leave democracy and humanity defenseless. This is why I suggest setting up multiple government-funded non-profit Frontier AI labs, distributed across several liberal democracies, and sharing of information across those labs so that in case one of the AGIs becomes rogue (in bad hands or out-of-control), the others can still help defend democracy and humanity. These labs would work on AI alignment and countermeasures against the misuse of AI or runaway AIs. Doing this safely and avoiding abuses of power calls for democratic governance which goes beyond the regulator, for example with a board that includes civil society and independent academics and representatives of other countries or of the international community.