Written Testimony 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 Judiciary
Subcommittee on Privacy, Technology, and the Law

July 25, 2023

EXECUTIVE SUMMARY

The capabilities of AI systems have steadily increased over the last two decades, often in surprising ways, thanks to the development of deep learning, for which I received the 2018 Turing Award with my colleagues Hinton and LeCun. These advancements have led many top AI researchers, including us three, to revise our estimates of when human levels of broad cognitive competence will be achieved. Previously thought to be decades or even centuries away, I and other leading AI scientists now believe human-level AI could be developed within the next two decades, and possibly within the next few years. The nature of digital computers compared to biological hardware suggests that such capability levels might then give AI systems significant intellectual advantages over humans.

Progress in AI has opened exciting opportunities for numerous beneficial applications that have driven researchers like myself throughout our careers. These advancements have rightfully attracted significant industrial investments and allowed rapid progress, for example in computer vision, natural language processing and molecular modeling. However, they also introduce new negative impacts and risks against which comparatively little investment has been made. These risks are challenging to assess, yet some have the potential to be catastrophic on a global scale. These range from major threats to democracy and national security, to the possibility of creating new entities more capable than humans, with potential loss of control over the course of humankind's future.

In the following sections, I will explain how such catastrophic outcomes could arise, emphasizing four factors that governments can influence to reduce the probability of such events. These factors include: (1) access - who can tinker with powerful AIs, what protocols must they follow, under what kind of oversight; (2) misalignment - the challenge of ensuring that AIs will act as intended, mitigating the fallout if they don't, and banning powerful AI systems that are not convincingly safe; (3) raw intellectual power - the capabilities of an AI system, which depend on the sophistication of its underlying algorithms and the computing resources and datasets on which it was trained; and (4) scope of actions - the ability to affect the world and cause harm in spite of society's defenses.

Importantly, none of the current advanced AI systems are demonstrably safe against the risk of loss of control to a misaligned AI. To minimize this risk as well as others, I propose actions that governments can take by addressing the aforementioned four factors.

- First, the accelerated implementation of agile national and multilateral regulatory frameworks and legislation that prioritize safety of the public from *all current and anticipated risks and harms* associated with AI, with more severe risks requiring more scrutiny.

- Second, the significant increase in global research endeavors focused on AI safety and governance to understand existing and future risks better, as well as study possible mitigation measures, both technical and normative. This open-science research should concentrate on safeguarding human rights and democracy, enabling the informed creation of essential regulations, safety protocols, safe AI methodologies, and governance structures.

- Third, investing now in research and development of shared as well as classified countermeasures to protect citizens and society from potential rogue AIs or AI-equipped bad actors with harmful goals. This work should be conducted within several highly secure and decentralized laboratories operating under multilateral oversight, aiming to minimize the risks associated with an AI arms race among governments or corporations.

The magnitude of these risks is so considerable that we should mobilize our best minds and ensure major investments in these efforts, on par with past efforts such as the space program or nuclear technologies - in order to fully reap the economic and social benefits of AI, while protecting societies, humanity and our shared future.

And, in the face of rapid technological change and the growing ubiquity of AI in society, there is an urgent need for policy action. We cannot afford to wait until a crisis - or “Black swan” event (low probability, high impact) occurs to react. The never before seen pace of development, deployment and adoption requires immediate, proactive and deliberate measures. Without such rapid adoption of governance mechanisms, I believe there are significant chances that the risks AI poses will far outweigh the innovation opportunities it may otherwise enable.

STRONG CONVICTIONS ON AI RESEARCH AND DEVELOPMENT

From the beginning of my graduate studies in the 80s, I made a deliberate choice to embark on research concerning artificial neural networks, which later gave rise to the advent of deep learning in the 2000s. I was motivated by an innate curiosity to comprehend the essence of intelligence, both within the natural world and in our capacity to craft artificial intelligences. The approach I pursued, centered around learning abilities and brain-inspired computation, was driven by the hypothesis that there exist scientific principles capable of elucidating the nature of intelligence, analogous to the fundamental principles that underpin the entirety of physics. The remarkable progress witnessed over the past two decades in the realms of deep learning and modern AI serves as compelling evidence that this is indeed the case.

In the 2010s, another motivating factor for my research emerged: the potential of AI to benefit humanity in numerous ways. For several years, AI has been driving a new scientific and economic revolution: from helping us discover new medications, to improving our ability to address pandemics, to providing new tools to fight the climate crisis, all while improving efficiency and productivity across many sectors of the economy. As a university professor leading a sizable research group, I considered it my responsibility to invest a significant portion of my work in AI applications that may not receive adequate private investments. Examples of such areas include research on infectious diseases or the development of new technologies that can model and combat climate change. Just as governments invested in areas such as medical research, environmental research, military research, the space program and the early days of Silicon Valley, with greater public investment and attention, "AI for good" applications could yield exceptional benefits to society across many domains.

The increased use of AI has come with downsides too, and as such, I have dedicated considerable personal effort to raising awareness of possible negative impacts, such as human rights issues including race and gender discrimination, as well as AI-enabled weapons and emerging concentration of capacity/power at odds with democracy and market efficiency. Additionally, I have actively participated in the development of social norms, standards, and regulations at both national and international levels. Notably, my work includes contributions to initiatives like the Montreal Declaration for a Responsible Development of AI, the Global Partnership on Artificial Intelligence (linked to the OECD), and serving on the Advisory Council on Artificial Intelligence for the Government of Canada. These endeavors aim to ensure that AI progresses in a responsible and ethically aligned manner.

GENERATIVE AI: THE TURNING POINT

Recent years have seen impressive advancements in the capabilities of generative AI, starting with image, speech, and video generation, more recently extended to natural language and made available to the public with OpenAI's ChatGPT, Microsoft's Bing Chat, Google's Bard and Anthropic's Claude. As a consequence, many AI researchers, including myself, have significantly revised our estimates regarding the timeline for achieving human-level AI systems, i.e., comparable to or stronger than humans on most cognitive tasks. Previously, I had placed a plausible timeframe for this achievement somewhere between a few decades and a century. However, along with my esteemed colleagues and co-recipients of the Turing Award for deep learning, Geoff Hinton and Yann LeCun, I now believe this plausible timeframe is within a few years to a couple of decades. The shorter timeframe, say within 5 years, is particularly worrisome because scientists, regulators and international organizations will most likely require a significantly longer timeframe to effectively mitigate the potentially significant threats to democracy, national security and our collective future.

While the scientific methodology behind these systems was not in itself revolutionary, the massive capability increase that comes from combining this methodology with large-scale training data and computational resources to train the AI was indeed unexpected and concerning for me and many others. This qualitative improvement caught many experts like myself off-guard and represented an unprecedented moment in history. Essentially, scientific progress has now reached what the computing pioneer Alan Turing proposed in 1950 as a milestone of future AI capability—the point at which it becomes challenging to discern in a text chat whether one is interacting with another human or a machine, commonly known as the Turing test. The current version of ChatGPT can feel human to many of us, indicating that there are now AI systems capable of mastering at least surface-level language and possessing sufficient knowledge about humankind to engage in highly proficient and creative, if sometimes unreliable, discussions. The next versions of this product will doubtless show significant improvements and make fewer mistakes. That is not to say that human-level AI has been reached. Whereas Geoff Hinton believes that the necessary ingredients are likely already known, Yann LeCun and myself believe that we have mostly figured out the principles giving rise to intuitive intelligence, but we are still missing aspects of cognition related to reasoning. Yet, my own work in this space leads me to believe that AI researchers could be close to a breakthrough on these missing pieces.

Contemplating the numerous instances in the past decade when the pace of AI advancements surpassed expectations, one must ponder where we are headed and what the implications might be, both positive and negative. Several factors suggest that once we can develop AI systems based on principles akin to those underlying human intelligence, these systems will likely surpass human intelligence in most cognitive tasks, i.e., we will have superhuman AIs. This notion was emphasized by Geoff Hinton in a recent conference, where he argued that, because AI systems are running on digital computers, they enjoy significant advantages over human brains. For instance, they can learn extremely fast by simultaneously consuming multiple sources of data across connected computers, which explains how ChatGPT was able to absorb a substantial fraction of Internet texts in just a few months, a feat that would require tens of thousands of human lives even if an individual were to spend every day reading. Additionally, AI systems can last virtually indefinitely, their programs and internal states can be easily replicated and copied across computers, akin to computer viruses, while our very mortal human brains are constrained by our continuously aging bodies.

THE DECOUPLING OF COGNITIVE ABILITIES FROM VALUES AND GOALS

To better understand the potential threats from these AI systems, we highlight here an important technical challenge faced by researchers when designing AI systems capable of effectively addressing cognitive tasks in a beneficial manner. This challenge arises from a critical distinction and separation between (a) desired outcomes, specified by goals and values, and (b) the efficient means of achieving those outcomes, relying on the cognitive abilities required to solve problems. Importantly, progress in AI can be achieved by separately (a) defining goals that align well with our desired results and underlying values and (b) determining optimal strategies for achieving these goals. This separation draws a parallel to the realm of economics, where a distinction exists between (a) the content of a contract (the goals), wherein Company A entrusts Company B with delivering specific outcomes, and (b) Company B's competence in achieving those goals.

Let us consider this decoupling between goals and cognitive competence in the case of an AI in the hands of a bad actor. In AI systems, it is relatively easy to replace a beneficial goal, such as summarizing a report, with a malicious one, such as generating disinformation, by modifying its instructions. A capable natural language interface implies that even non-experts may be able to introduce malevolent goals, as illustrated recently in the case of GPT-4 being coaxed by non-experts to provide advice to design pandemic-grade pathogens or to find cybersecurity vulnerabilities. Furthermore, as illustrated with AutoGPT, it is fairly easy to turn a question-answering system like ChatGPT into a system that can take action on the internet, without a human in the loop - which greatly increases the potential for harm.

Let us now consider the case of someone with no malicious intent operating a powerful AI system. Much progress has been made in recent years regarding the development of cognitive abilities to perform tasks specified by given goals, but we still have no way to guarantee that the AI systems will perform as we intend when specifying those goals. This problem is not unique to AI: it was the subject of the 2016 Nobel Prize in Economics, and is relatable to any lawmaker who has witnessed citizens or corporations subverting the spirit of the law while following the letter of the law. In a contract between two parties, it is impractical for Party A to fully specify Party B's responsibilities, because it requires enumerating every possible circumstance in the contract. This makes it possible for Party B to adhere to the letter of the contract while exploiting loopholes that leave the spirit of the contract unfulfilled. In AI, the act of designing a goal is very much like writing a contract, and the challenge of specifying goals with intended effects is known as the alignment problem, which is unsolved. Just as Party B might understand the spirit of the contract, but still stick to the letter of it, an AI that is misaligned with its designers would not “correct” its behavior. This misalignment already manifests in the present harms caused by AI systems, such as when a dialogue system insults a user, or when an AI company unintentionally designs a computer vision system with significantly poorer performance in recognizing the faces of Black individuals.

As AI systems increasingly surpass human intelligence in various domains, the concern arises whether these misalignments could result in more substantial and widespread harm, whether directed by a human or not. Consequently, proactive consideration of policies that can mitigate such risks before they materialize becomes imperative.

HOW AI MAY CAUSE MAJOR HARMS

Let us consider some of the main scenarios that worry me particularly because they could yield major harms by superhuman AIs.

(1) The first is the **use of an AI system as an intentionally harmful tool.** This is already a possibility with present systems, and would be enhanced by future algorithms with superhuman capabilities. Current and upcoming AI systems are likely to lower the barrier to entry for dual-use research and technology on both the beneficial and dangerous sides, making powerful tools readily accessible to more people. For example, an AI developed with data from molecular biology can be used to design medicines, but can also be used to design a bioweapon or chemical weapon requested by a bad actor. The same would go for the design of computer viruses that could defeat our current cybersecurity defenses. While these actions were possible prior to AI, the degree to which they are facilitated and semi-automated by AI means that a much broader swath of non-experts and malicious actors would now have these capabilities at their disposal. The risks proliferate when humans are not required to be in the loop - for example, if an algorithm is given free access to social media and can coordinate large-scale disinformation campaigns. The more extreme future case would be when an AI system is autonomous, i.e., when it can perform actions directly, for example order DNA on the internet from biotechnology companies and hire humans (who might not realize their role) as part of a scheme to assemble the different pieces of the puzzle that corresponds to a highly lethal and virulent pathogen.

(2) In the second scenario, **unintended harm is inflicted by an AI system used as a tool** - for example, if it fails in rare circumstances, or involves subtle biases that lead to consistently lower performance for certain users. This kind of situation occurs frequently now, for example when an AI algorithm for granting loans is biased against people of color, because the data it was trained on was biased and/or the teams designing them did not adequately consider demographic biases in the design of the algorithm itself. Another example would be the interface between AI and military weapon systems where the propensity of human operators to follow the fallible recommendation of computers, combined with a subtly misaligned system, could yield grave consequences in a nuclear threat scenario.

(3) The third possibility, which could emerge in as little as a few years, is that of **loss of control**, when an AI is given a goal that includes or implies maintenance of its own agency, which is equivalent to a survival objective. This can be intentional by the human creator, or may arise implicitly as a means to achieve a human-given goal (in a manner reminiscent of the movie 2001: A Space Odyssey). Indeed, an AI system may conclude that in order to achieve the given goal, it must not be turned off. If a human then tries to turn it off, a conflict may ensue. This may sound like science fiction, but it is sound and real computer science. We run into the alignment challenge described above: it is difficult to perfectly specify all of our expectations of the AI behavior. This misalignment opens the door to harm that can become catastrophic as AI systems become more and more capable, because loopholes tend only to be fixed after they have been exploited. One may believe that we could fix the original human-specified goal to avoid harmful misalignment, filling in edge cases that we omitted, but we are not likely to be able to patch every omission one by one without incurring potentially major or irreparable harm at each step. If the AI is misspecified, powerful enough, and exploits a loophole in its goals, the consequences could be unforeseen and severe. Therefore, a reactive approach to mitigating misspecified goals could be extremely costly for society, and we may only have a few chances of getting the alignment right for superhuman AI.

Other scenarios have been discussed in the AI safety literature, but I am most concerned by the above. In the last few months, I have discussed these with many of my fellow AI researchers and considered both arguments in favor of lower levels of concern, as well as those that suggest we should on the contrary use extreme caution. I have listed these in an FAQ document about catastrophic AI risks on my personal blog. Although I acknowledge there exists a lot of uncertainty about the most extreme risks, the amplitude of potential negative impacts is such that I lean towards prudence, setting up preventative measures and investing massively in research to help shape a positive path forward.

One of the most relevant points raised in ongoing debates revolves around the question of how an AI system—a piece of code running on a computer—can inflict tangible harm in the physical world. While artificial systems have been around for decades, what is new now is that their level of “common sense” has risen enough to allow them to operate in the unconstrained real world. Let's consider illustrative scenarios where a computer equipped with superhuman AI capabilities, including superhuman programming and cybersecurity skills, is granted internet access and provided with a bank account. Would it be impossible for such an AI to infiltrate other computers and replicate itself across multiple locations to minimize the risk of being shut down? Would it be impossible for it to perform frauds and generally earn money online, for example through phishing or financial trading? Would it be impossible for it to influence humans or pay them to perform certain tasks or even recruit organized crime networks for illicit activities? With its cybersecurity expertise and the power to influence social media discussions and human decision-makers, couldn't a superhuman AI manipulate elections and the media, thus jeopardizing our democracies? With publicly available knowledge of biology and chemistry, couldn't a superhuman AI design bioweapons or chemical weapons? It is hard to have strong guarantees of the above impossibilities required for safety, once we consider the premise of superhuman AI capabilities.

In all cases, human involvement plays a critical role in enabling such harm, intentionally or not, through R&D efforts, insufficient understanding of consequences, lack of prudence / negligence, or as a subject of influence of the AI system. Government intervention and regulation that influences human behavior to achieve greater safety is thus essential.

In the long run, once systems that surpass humans in intelligence and possess sufficient power to cause harm (through human actors or directly) are created, it could potentially threaten the security of citizens across the globe and significantly disempower humanity. Given the great uncertainties surrounding the future beyond the advent of superhuman AI with considerable agency powers, it is imperative to consider every measure to avert such outcomes.

CONDITIONS FOR MAJOR HARM AS CHOKE POINTS TO MINIMIZE RISKS

For an AI to cause major harm, some conditions are required. They can be grouped into four categories in order to clarify the choke points where public policies could mitigate these risks:

(1) **Access**: **Limiting who and how many people and organizations have access to powerful AI systems, structuring the proper protocols, duties, oversight and incentives for them to act safely.** For example, very few people in the world are allowed to fly passenger jets or have a national security clearance, and they are selected based on required trustworthiness, skills and ethical integrity, which considerably reduces the chance of accidents. What sort of procedures do the designers/owners of these AI systems have to follow, and what incentives (including liability and regulations) do they have to act with care and ensure they do not cause harm? And how do we regulate access while avoiding concentration of power, e.g., in the hands of a few unelected individuals and/or large profit-driven companies?

(2) **Misalignment**: **Ensuring that AI systems will act appropriately, as intended by their operators and in agreement with our values and norms, mitigating against the potentially harmful impact of misalignment and banning powerful AI systems that are not convincingly safe.** What are the system's goals (programmed or developed),

how aligned are they with societal values, and how and by whom are these values legitimately established? How do we design tests to verify the quality of the alignment (e.g., with independent audits)? Could this misalignment cause significant harm with sufficient cognitive power and ability of the AI to act?

(3) **Raw intellectual power: Considering the ability of an AI system to understand the world and elaborate action plans, which depends on the level of sophistication of its algorithms** (mathematical principles and formulae designed by AI researchers or invented by the AI itself) **as well as the amount of compute and the diversity of data it uses for learning or sensing the world** (e.g. searching the web). How competent is the AI at actually understanding the world - or some aspects of it over which its actions could become dangerous - and at devising plans to achieve its goals? This suggests monitoring and possible restrictions of these sources of raw intellectual power, namely advanced algorithms, large computing capabilities and large/sensitive datasets.

(4) **Scope of actions: Evaluating the ability of the AI to influence individuals, affect the world, and cause harm indirectly** (e.g. through human actions) **or directly** (e.g. through the internet), **as well as society's ability to prevent or limit such harm.** What is the severity and scale of the harm these actions could cause? For example, an AI system that controls powerful weapons can do much more damage than one that only controls the heating and air conditioning of a building.

There is uncertainty surrounding the rate at which AI capabilities will increase. However, there is a significant probability that superhuman AI is just a few years away, outpacing our ability to comprehend the various risks and establish sufficient guardrails, particularly against the more catastrophic scenarios. The current "gold rush" into generative AI might, in fact, accelerate these advances in capabilities. Additionally, the far-reaching developments of the Internet, digital integration, and social media may amplify the scope of harm caused by such future advanced AI, especially rogue superhuman AI. We cannot afford to wait until a "Black swan" event (low probability, high impact, cascading effects and major disruptions) occurs to take action, as the pace of technological change means that we must be proactive. The COVID pandemic was an example of how rapid developments can catch us off guard, and how the need for preparedness and resilience is crucial. Consequently, it is urgent for governments to intervene with regulation and invest in research to protect our society, and I offer a suggested path forward below.

THE PATH FORWARD: REGULATING AI AND INVESTING IN RESEARCH

While there remains much to be understood about the potential for harm of very powerful AI systems, looking at risks through the lens of each of the above-mentioned four factors is critical to designing appropriate actions.

In light of the significant challenges societies face in designing the needed regulation and international treaties, I firmly believe that urgent efforts in the following areas are crucial:

a) **The coordination and implementation of agile national and multilateral regulations - beyond voluntary guidelines - anchored in new international institutions that prioritize public safety in relation to all risks and harms associated with AI.** This necessitates clear and mandatory, but evolving, standards for the comprehensive evaluation of potential harm through independent audits and restricting/prohibiting (with criminal law) the development and deployment of AI systems possessing dangerous capabilities. The goal should be to establish a level of scrutiny beyond that applied in the pharmaceutical, transportation, or nuclear industries. Minimal global standards should be set globally and enforced by domestic regulators, using the pressure of commercial barriers to maximize compliance with standards across the world.

b) **Significantly accelerating global research endeavors focused on AI safety and governance to enhance our comprehension of existing and future risks.** This research should be open-access and concentrate on safeguarding human rights and democracy, enabling the informed creation of essential regulations, safety protocols, safe AI methodologies, and new governance structures.

c) **Immediate investments in research and development aiming at designing countermeasures to minimize harm from potential rogue AIs, with paramount emphasis on safety.** This work should be conducted within highly secure and decentralized laboratories operating under multilateral oversight, in order to minimize the risks associated with an AI arms race or direct control by malicious actors or governments. A centralized research center would likely not be as efficient as a network of laboratories with independent and diverse research directions, and implementing these labs in several countries would make the network more robust. Neutral and autonomous entities that are ideally non-profit and non-governmental should lead this research, combining expertise in national and international security and AI, to ensure this work is uncompromised by national or commercial interests. They could be audited following safety rules set by the international community and participating governments, with an agreed upon mission to which products of work must align.

As expressed by Kelsey Piper regarding catastrophic risks of AI: "when there is this much uncertainty, high-stakes decisions shouldn’t be made unilaterally by whoever gets there first. If there were this much expert disagreement about whether a plane would land safely, it wouldn’t be allowed to take off — and that’s with 200 people on board, not 8 billion."

Given the significant potential for large-scale harm, governments must allocate substantial additional social and technological resources to safeguard our future, inspired by efforts such as space exploration or nuclear fusion. The UK AI task force is a good example of how to initiate such a movement and start acting now. As for regulatory frameworks, they should be extremely agile in order to quickly react to changes in technology, new research on safety and fairness, and nefarious uses that emerge. An example of such a framework is Canada's principle-based approach (The Artificial Intelligence and Data Act or AIDA), in which the law itself contains high-level objectives which are in turn defined, adapted and operationalized in regulation. This honors the important and necessary processes that lead to the adoption of laws, while providing agility for governmental bodies to design and adapt regulation as needed, thus keeping pace with technological developments.

ADDITIONAL THOUGHTS ON REGULATORY ACTION

While these regulatory and research efforts will unfold over the course of multiple years, a number of elements are already coming into focus that can/should be enacted, namely regarding access, monitoring and evaluating potential for harm. Additional thoughts on appropriate policies (as per the four choke points above), include:

- Ethics review committees or boards in academic and industrial labs developing algorithms or trained models that could bring rapid advances in AI capabilities;
- Requiring documentation of the development process and the safety analysis of AI systems over multiple stages - before training, before deployment, and ongoing - to enable auditing and verification of safety protocols;
- Ensuring that AI-generated content is identified as such to users to reduce the influence of AI systems (controlled by malicious individuals or not) on people's opinions to minimize the risk that people mistakenly believe AI-generated content to be real;
- Licenses for companies and people with access to highly capable systems, monitoring of advanced AI systems, and who works with them, ensuring conformity to established risk-minimizing procedures;
- Registration requirements for advanced AIs trained with more than a specified amount of compute;
- Keeping track of the size and scope of the datasets used to train systems to differentiate AI systems that are highly specialized (targeted field of action) from those that are very general-purpose and can interact with / influence / manipulate citizens and society;
- Limiting access to source code and trained advanced models (beyond a critical threshold of competency) to individuals and organizations with the appropriate licensing. Furthermore, to avoid concentration of power in the hands of a few licensed corporations, a substantial fraction of these licensed organizations should be bound to spread the benefits, through public funding and/or global public good objectives;
- Strict regulatory requirements or bans on the development of highly advanced AIs known for the risk of emergent goals within an AI, such as reinforcement learning, until we have clear evidence of their safety;
- Semi-automated screening of powerful AI systems for requests that can lead to dangerous behaviors such as terrorism or to increasing the power of the AI;
- Controlling and limiting the ability of highly capable AI systems to act in the world (for example via the Internet or specialized tools);

- Associating social media and email accounts with a well-identified human being who registered in person with an ID, making it harder for AI systems to rapidly take over a large number of social media or email accounts;
- Monitoring and restriction of biotechnology and pharmaceutical companies' sharing of sensitive data and creation of new or genetically modified biological organisms (that could be used for bioweapons).

Since the Internet and social media have no strong national borders, nor do biological or computer viruses, it will of course be critically important to negotiate international agreements such that public policies and regulations aiming at reducing the risks of catastrophic outcomes from AI are well synchronized worldwide. An international treaty and supporting UN agency akin to the IAEA are necessary to standardize access permissions, cybersecurity countermeasures, safety restrictions and fairness requirements of AI globally. The world has widely varying cultures and norms, making agreed upon principles such as the UN Universal Declaration of Human rights a good base from which to expand. However, safety against rogue AIs, with the future of all of humanity at stake, suggests we aim for a worldwide treaty on AI safety, AI governance and countermeasures.

CONCLUSION

As expressed through this testimony, I am very concerned by the severe and potentially catastrophic risks that could arise intentionally - because of malicious actors using advanced AI systems to achieve harmful goals, or unintentionally - if an AI system develops strategies to achieve its objectives that are misaligned with our values. I am grateful to have had the opportunity to present my perspective, emphasizing four factors that governments can focus on in their regulatory efforts to mitigate harms, especially major ones, associated with AI.

I feel strongly that it is critical to invest immediately and massively in research endeavors to design systems and safety protocols that will minimize the probability of yielding rogue AIs, as well as develop countermeasures against the possibility of undesirable scenarios. There is a great need and opportunity for innovation in governance research to design adaptable and agile regulations and treaties that will safeguard citizens and society as the technology evolves and/or new unexpected threats arise.

I believe we have the moral responsibility to mobilize our greatest minds and major resources in a bold coordinated effort to fully reap the economic and social benefits of AI, while protecting society, humanity and our shared future against its potential perils. And we need to do so urgently, with the U.S. playing the same leadership role in protecting humanity as it is in advancing AI capabilities.

BIOGRAPHY

Yoshua Bengio is recognized worldwide as one of the leading experts in artificial intelligence, known for his conceptual and engineering breakthroughs in artificial neural networks and deep learning. He is a Full Professor in the Department of Computer Science and Operations Research at Université de Montréal and the Founder and Scientific Director of Mila – Quebec Artificial Intelligence Institute, one of the largest academic institutes in deep learning and one of the three federally-funded centers of excellence in AI research and innovation in Canada.

He obtained his Ph.D. in Computer Science from McGill University in 1991, in Montreal. After completing a postdoctoral fellowship in 1991-1992 at the Massachusetts Institute of Technology (MIT) on statistical learning and sequential data modeling, he completed a second postdoc at AT&T Bell Laboratories, in Holmdel, NJ, on learning and vision algorithms in 1992-1993. In September 1993, he returned to Montreal and joined U. Montreal as a faculty member.

In 2016, he became the Scientific Director of the IVADO institute and obtained the largest grant in the university's history (94M$CAN). He is Co-Director of the CIFAR Learning in Machines & Brains program that funded the initial breakthroughs in deep learning and since 2019, holds a Canada CIFAR AI Chair and is Co-Chair of Canada's Advisory Council on AI.

In 2022, Yoshua Bengio became the most cited computer scientist in the world in terms of h-index. Both motivated by the growth of the AI startup ecosystem and concerned about the social impact of AI since the industrialization of AI in 2013, he actively took part in the conception of the Montreal Declaration for the Responsible Development of Artificial Intelligence. His goal is to contribute to uncovering the principles giving rise to intelligence through learning while favouring the safe development of AI for the benefit of all.

Yoshua Bengio was made an Officer of the Order of Canada and a Fellow of the Royal Society of Canada in 2017 and in 2020, became a Fellow of the Royal Society of London. From 2000 to 2019, he held the Canada Research Chair in Statistical Learning Algorithms. He is a member of the NeurIPS Foundation advisory board and Co-Founder of the ICLR conference.

His scientific contributions have earned him numerous awards, including the 2019 Killam Prize for Natural Sciences, the 2017 Government of Québec Marie-Victorin Award, the 2018 Lifetime Achievement Award from the Canadian AI Association, the Prix d'excellence FRQNT (2019), the Medal of the 50th Anniversary of the Ministry of International Relations and Francophonie (2018), the 2019 IEEE CIS Neural Networks Pioneer Award, Acfas's Urgel-Archambault Prize (2009) and in 2017, he was named Radio-Canada's Scientist of the Year.

He is the 2018 laureate of the A.M. Turing Award, "the Nobel Prize of Computing," alongside Geoffrey Hinton and Yann LeCun for their important contributions and advances in deep learning. In 2022, he was appointed Knight of the Legion of Honor by France and named co-laureate of Spain's Princess of Asturias Award for technical and scientific research.

ACKNOWLEDGMENTS

This document benefited from the feedback of Valérie Pisano, Niki Howe, Michael Cohen, David Rolnick, Alan Chan, Richard Mallah, Benjamin Prudhomme, Julia Bossmann, Sören Mindermann, Lama Saouma, Marc-Antoine Guérard, Dan Hendrycks, Noam Kolt, Roger Grosse, Ludovic Soucisse, Alex Hernandez-Garcia, Cristian Dragos Manta, Edward J. Hu, Fazl Barez, Jean-Pierre Falet.